Percona XtraDB Cluster:

Failure Scenarios and their Recovery

Krunal Bauskar (PXC Lead, Percona)
Alkin Tezuysal (Sr. Technical Manager, Percona)

Who we are?

Krunal Bauskar

- Database enthusiast.
- Practicing databases (MySQL) for over a decade now.
- Wide interest in data handling and management.
- Worked on some real big data that powered application @ Yahoo, Oracle, Teradata.

Alkin Tezuysal (@ask_dba)

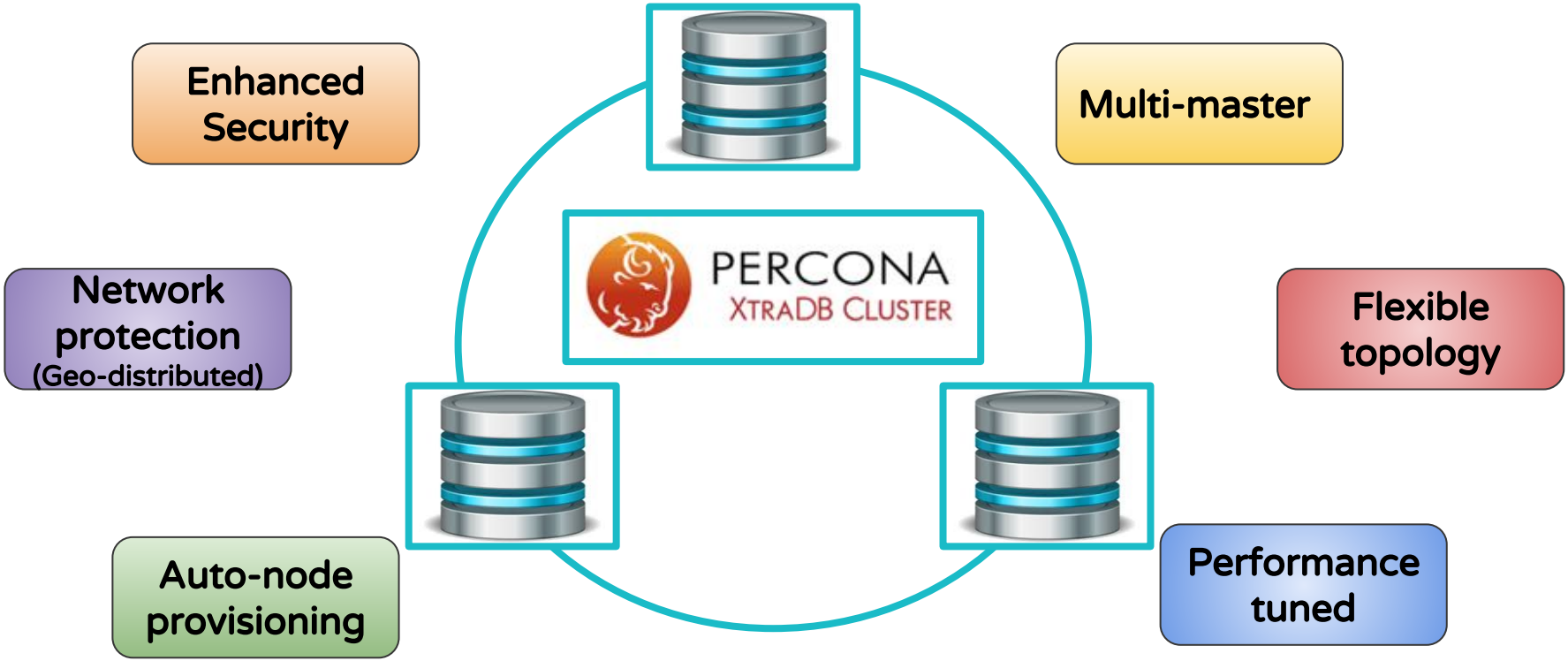
- Open Source Database Evangelist
- Global Database Operations Expert
- Cloud Infrastructure Architect AWS
- Inspiring Technical and Strategic Leader
- Creative Team Builder
- Speaker, Mentor, and Coach
- Outdoor Enthusiast

Agenda

- Quick sniff at PXC
- Failure Scenarios and their recovery
- PXC Genie - You wish. We implement.
- Q & A

Quick Sniff at PXC

What is PXC ?


Enhanced Security
Multi-master
Network protection (Geo-distributed)
Flexible topology
Auto-node provisioning
Performance tuned
PERCONA
XTRADB CLUSTER

Failure Scenarios and their recovery

Scenario: New node fail to connect to cluster

Scenario: New node fail to connect to cluster

```
2018-10-12T05:35:41.532125Z 0 [Note] WSREP: gcomm: connecting to group 'pxc-cluster', peer '192.168.10.1:'
2018-10-12T05:35:44.629768Z 0 [Note] WSREP: announce period timed out (pc.announce_timeout)
2018-10-12T05:35:44.630170Z 0 [Warning] WSREP: no nodes coming from prim view, prim not possible
2018-10-12T05:35:44.630247Z 0 [Note] WSREP: Current view of cluster as seen by this node
view (view_id(NON_PRIM,a8551297,1)
memb {
        a8551297,0
        }
joined {
        }
left {
        }
partitioned {
        }
)
2018-10-12T05:35:45.328882Z 0 [Warning] WSREP: last inactive check more than PT1.5S (3*evs.inactive_check_period) ago (PT3.79679S), skipping check
2018-10-12T05:36:14.703885Z 0 [Note] WSREP: Current view of cluster as seen by this node
view ((empty))
2018-10-12T05:36:14.704184Z 0 [ERROR] WSREP: failed to open gcomm backend connection: 110: failed to reach primary view (pc.wait_prim_timeout): 110 (Connection timed out)
         at gcomm/src/pc.cpp:connect():159
2018-10-12T05:36:14.704220Z 0 [ERROR] WSREP: gcs/src/gcs_core.cpp:gcs_core_open():209: Failed to open backend connection: -110 (Connection timed out)
2018-10-12T05:36:14.704968Z 0 [ERROR] WSREP: gcs/src/gcs.cpp:gcs_open():1514: Failed to open channel 'pxc-cluster' at 'gcomm://192.168.10.1': -110 (Connection timed out)
2018-10-12T05:36:14.705010Z 0 [ERROR] WSREP: gcs connect failed: Connection timed out
2018-10-12T05:36:14.705237Z 0 [ERROR] WSREP: Provider/Node (gcomm://192.168.10.1) failed to establish connection with cluster (reason: 7)
2018-10-12T05:36:14.705269Z 0 [ERROR] Aborting
```

Joiner log

Scenario: New node fail to connect to cluster

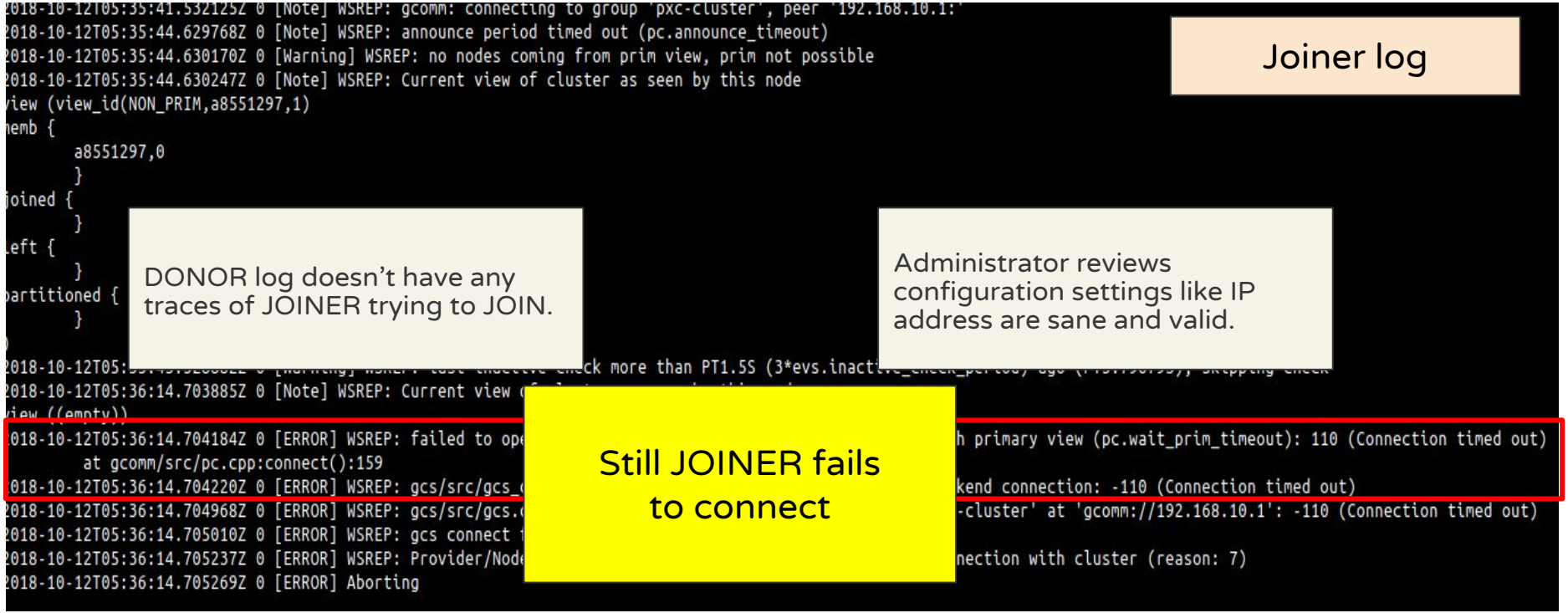
```
2018-10-12T05:35:41.532125Z 0 [Note] WSREP: gcomm: connecting to group 'pxc-cluster', peer '192.168.10.1:'
2018-10-12T05:35:44.629768Z 0 [Note] WSREP: announce period timed out (pc.announce_timeout)
2018-10-12T05:35:44.630170Z 0 [Warning] WSREP: no nodes coming from prim view, prim not possible
2018-10-12T05:35:44.630247Z 0 [Note] WSREP: Current view of cluster as seen by this node
view (view_id(NON_PRIM,a8551297,1)
memb {
        a8551297,0
        }
joined {
        }
left {
        }
partitioned {
        }
)
2018-10-12T05:                                                    ck more than PT1.5S (3*evs.inact
2018-10-12T05:36:14.703885Z 0 [Note] WSREP: Current view of cluster as seen by this node
view ((empty))
2018-10-12T05:36:14.704184Z 0 [ERROR] WSREP: failed to open gcomm backend connection: 110: failed to reach primary view (pc.wait_prim_timeout): 110 (Connection timed out)
         at gcomm/src/pc.cpp:connect():159
2018-10-12T05:36:14.704220Z 0 [ERROR] WSREP: gcs/src/gcs_core.cpp:gcs_core_open():209: Failed to open backend connection: -110 (Connection timed out)
2018-10-12T05:36:14.704968Z 0 [ERROR] WSREP: gcs/src/gcs.cpp:gcs_open():1514: Failed to open channel 'pxc-cluster' at 'gcomm://192.168.10.1': -110 (Connection timed out)
2018-10-12T05:36:14.705010Z 0 [ERROR] WSREP: gcs connect failed: Connection timed out
2018-10-12T05:36:14.705237Z 0 [ERROR] WSREP: Provider/Node (gcomm://192.168.10.1) failed to establish connection with cluster (reason: 7)
2018-10-12T05:36:14.705269Z 0 [ERROR] Aborting
```

Joiner log

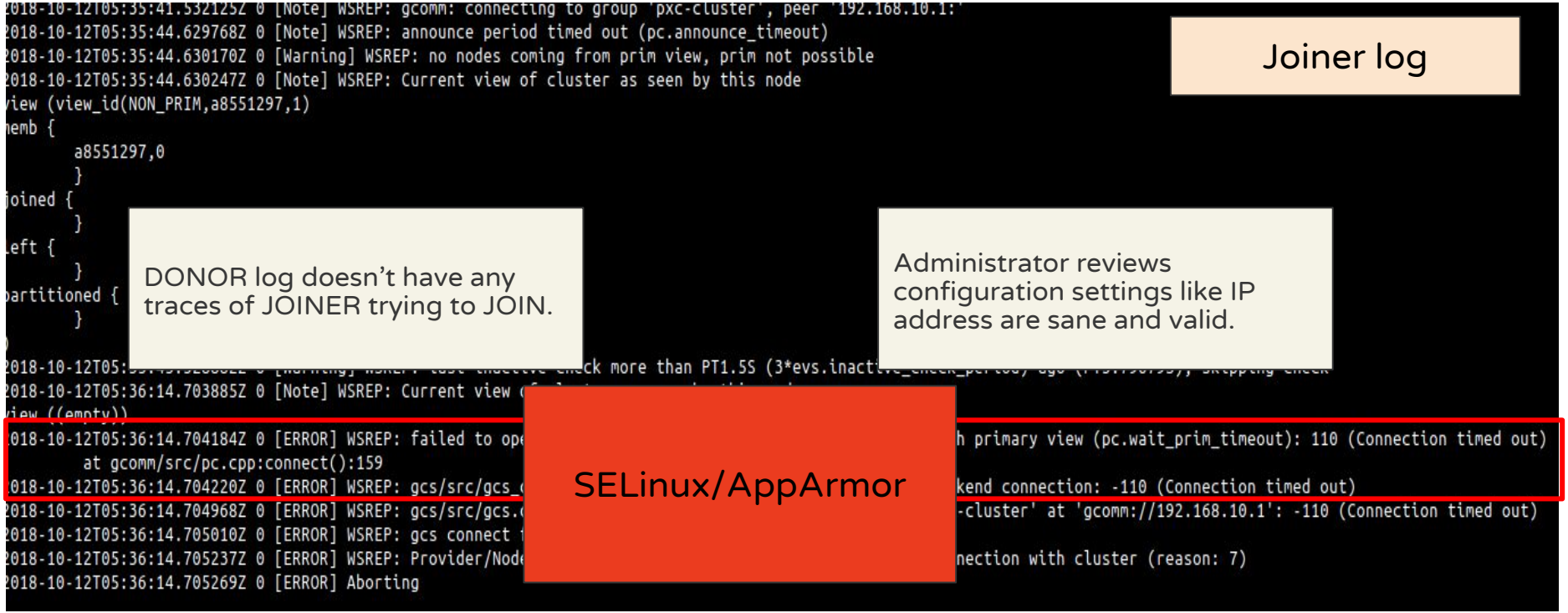
DONOR log doesn't have any traces of JOINER trying to JOIN.

Administrator reviews configuration settings like IP address are sane and valid.

Scenario: New node fail to connect to cluster



Scenario: New node fail to connect to cluster



Scenario: New node fail to connect to cluster

```
2018-10-12T05:35:41.532125Z 0 [Note] WSREP: gcomm: connecting to group 'pxc-cluster', peer '192.168.10.1:'
2018-10-12T05:35:44.629768Z 0 [Note] WSREP: announce period timed out (pc.announce_timeout)
2018-10-12T05:35:44.630170Z 0 [Warning] WSREP: no nodes coming from prim view, prim not possible
2018-10-12T05:35:44.630247Z 0 [Note] WSREP: Current view of cluster as seen by this node
view (view_id(NON_PRIM,a8551297,1)
memb {
        a8551297,0
        }
joined {
        }
left {
        }
partitioned {
        }
)
2018-10-12T05:35:45.328882Z 0 [Warning] WSREP: last inactive check more than PT1.
2018-10-12T05:36:14.703885Z 0 [Note] WSREP: Current view of cluster as seen by th
view ((empty))
2018-10-12T05:36:14.704184Z 0 [ERROR] WSREP: failed to open gcomm backend connection: 110: failed to reach primary view (pc.wait_prim_timeout): 110 (Connection timed out)
         at gcomm/src/pc.cpp:connect():159
2018-10-12T05:36:14.704220Z 0 [ERROR] WSREP: gcs/src/gcs_core.cpp:gcs_core_open():209: Failed to open backend connection: -110 (Connection timed out)
2018-10-12T05:36:14.704968Z 0 [ERROR] WSREP: gcs/src/gcs.cpp:gcs_open():1514: Failed to open channel 'pxc-cluster' at 'gcomm://192.168.10.1': -110 (Connection timed out)
2018-10-12T05:36:14.705010Z 0 [ERROR] WSREP: gcs connect failed: Connection timed out
2018-10-12T05:36:14.705237Z 0 [ERROR] WSREP: Provider/Node (gcomm://192.168.10.1) failed to establish connection with cluster (reason: 7)
2018-10-12T05:36:14.705269Z 0 [ERROR] Aborting
```

Joiner log

Don't confuse this error with SST since node is not yet offered membership of cluster. SST comes post membership.

Scenario: New node fail to connect to cluster

- Solution-1:
  - Setting mode to PERMISSIVE or DISABLED

Scenario: New node fail to connect to cluster

- Solution-1:
  - Setting mode to PERMISSIVE or DISABLED
- Solution-2:
  - Configuring policy to allow access in ENFORCING mode.
  - Related blogs
    - *"Lock Down: Enforcing SELinux with Percona XtraDB Cluster". It probs what all permission are needed and add rules accordingly.*
    - "Lock Down: Enforcing AppArmor with Percona XtraDB Cluster"
    - *Using this we can continue to use SELinux in enable mode. (You can also refer to selinux configuration on Codership site too).*

Scenario: New node fail to connect to cluster

PXC can operate with SELinux/AppArmor.

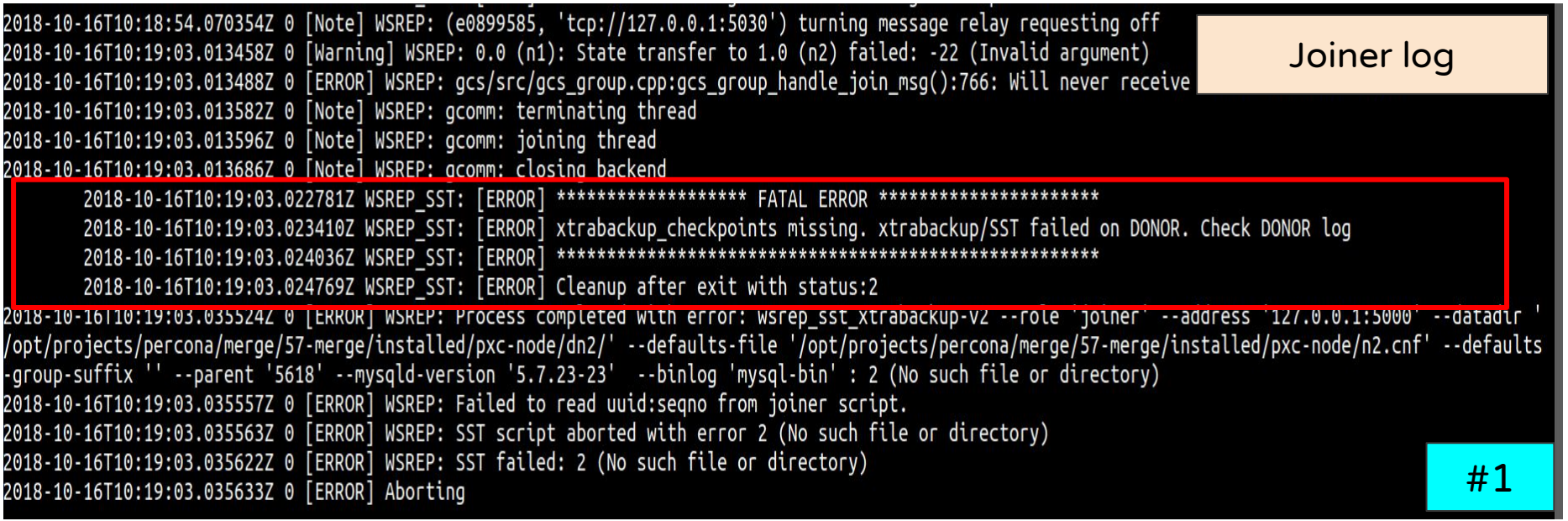
Scenario: Catching up cluster (SST, IST)

Scenario: Catching up cluster (SST, IST)

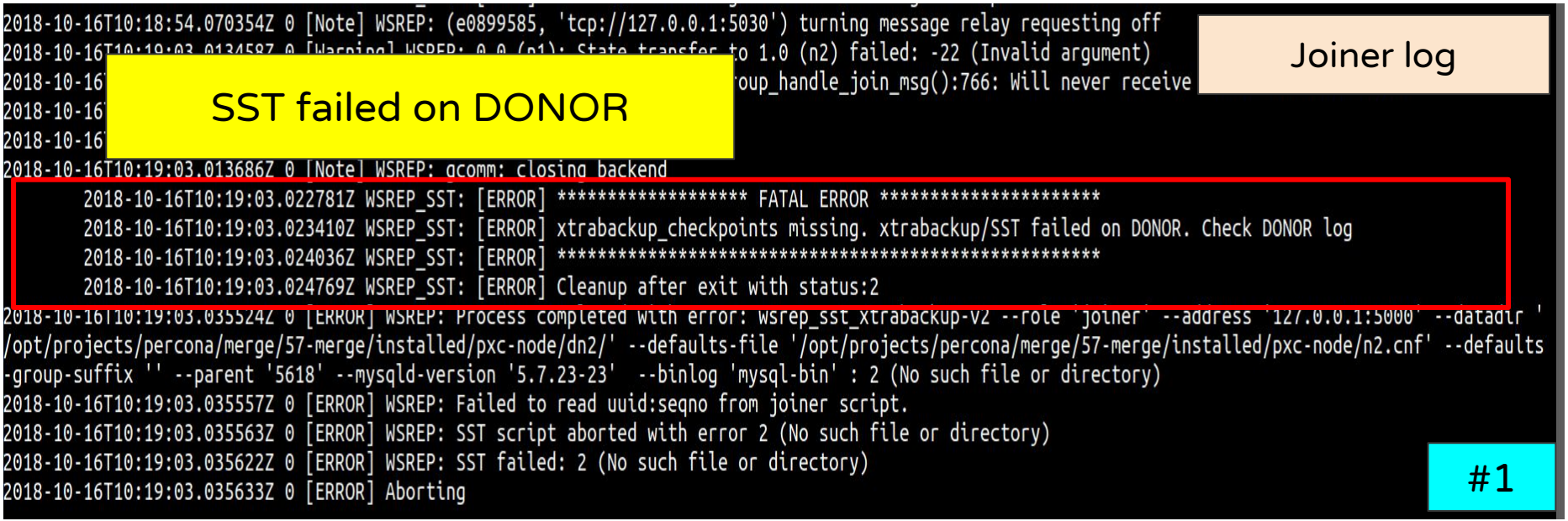
- SST: complete copy-over of data-directory
  - SST has has multiple external components SST script, XB, network aspect, etc. Some of these are outside control of PXC process.

- IST: missing write-sets (as node is already member of cluster).
  - Intrinsic to PXC process space.

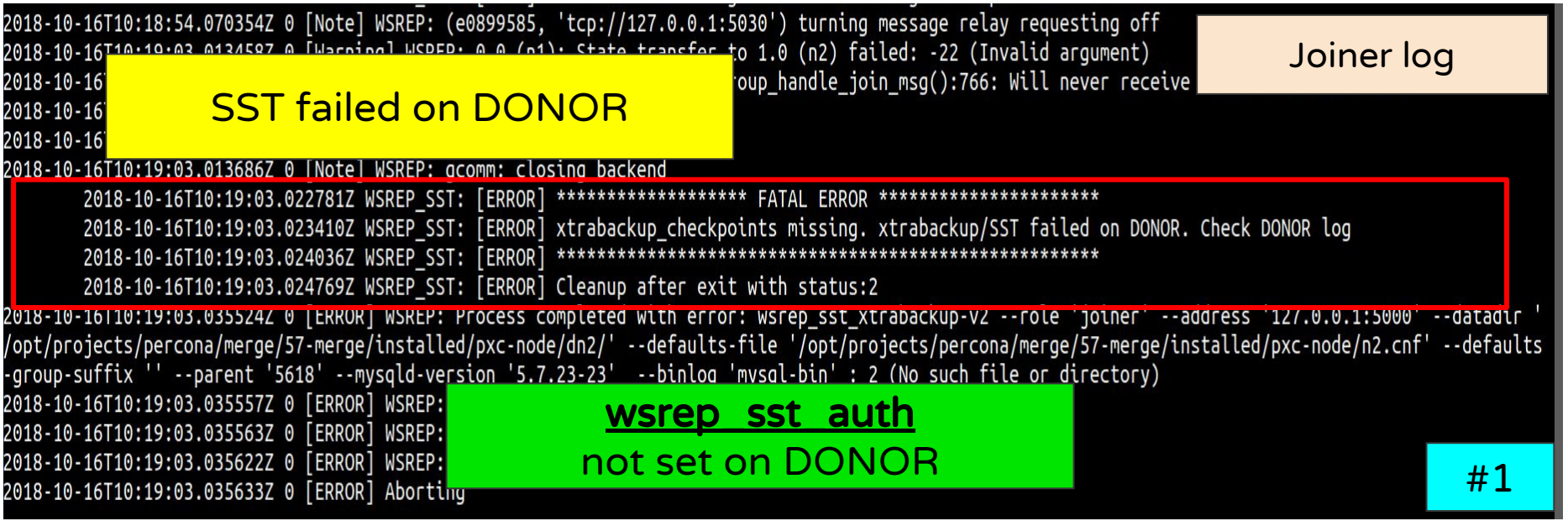
Scenario: Catching up cluster (SST, IST)



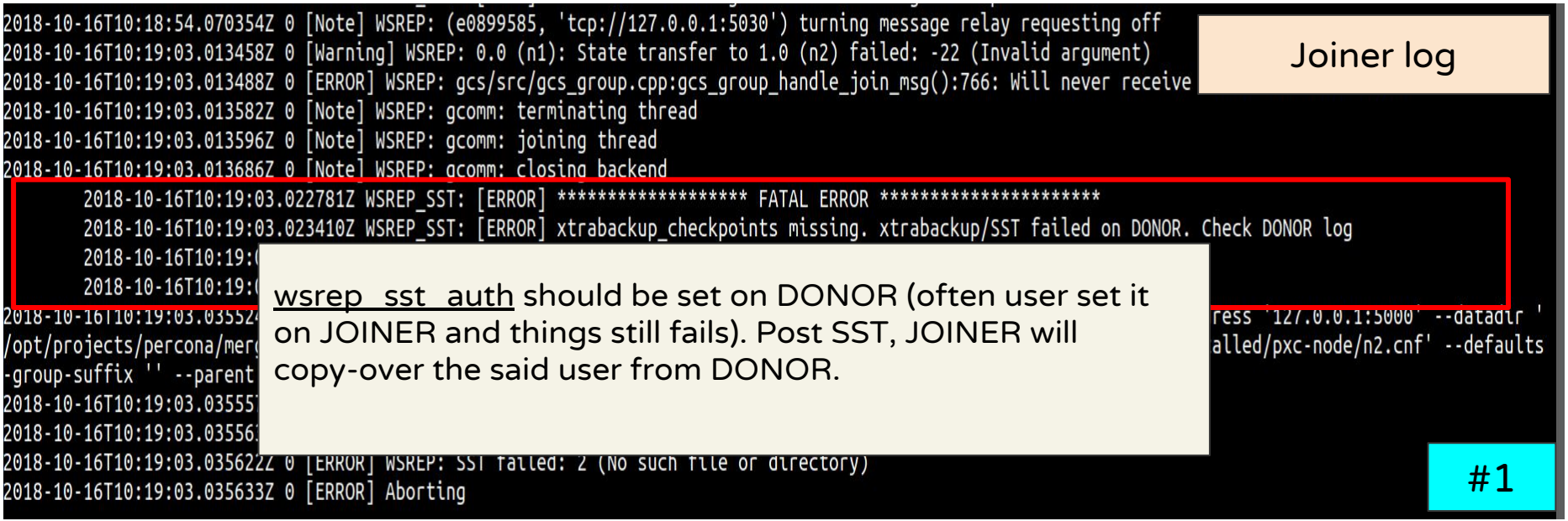
Scenario: Catching up cluster (SST, IST)



Scenario: Catching up cluster (SST, IST)



Scenario: Catching up cluster (SST, IST)



Scenario: Catching up cluster (SST, IST)

```
2018-10-16T10:19:02.998914Z WSREP_SST: [ERROR] ********** FATAL ERROR **********
2018-10-16T10:19:02.999567Z WSREP_SST: [ERROR] xtrabackup finished with error: 1.  Check /opt/projects/percona/
dn1//innobackup.backup.log
------------- innobackup.backup.log (START) ------------------
xtrabackup: recognized server arguments: --datadir=/opt/projects/percona/merge/57-merge/installed/pxc-node/dn1 --log_bin=mysql-bin --server-id=2 --defaul
ts_group=mysqld --parallel=4
xtrabackup: recognized client arguments: --datadir=/opt/projects/percona/merge/57-merge/installed/pxc-node/dn1 --log_bin=mysql-bin --server-id=2 --defaul
ts_group=mysqld --parallel=4 --socket=/tmp/n1.sock --lock-ddl=1 --backup=1 --galera-info=1 --binlog-info=ON --stream=xbstream --target-dir=/tmp/pxc_sst_h
5zR/donor_xb_SE1L
encryption: using gcrypt 1.8.1
181016 15:49:02  version_check Connecting to MySQL server with DSN 'dbi:mysql:;mysql_read_default_group=xtrabackup;mysql_socket=/tmp/n1.sock' (using pass
word: NO).
Failed to connect to MySQL server as DBD::mysql module is not installed at - line 1327.
181016 15:49:02 Connecting to MySQL server host: localhost, user: not set, password: not set, port: not set, socket: /tmp/n1.sock
Failed to connect to MySQL server: Access denied for user 'kbauskar'@'localhost' (using password: NO).
------------- innobackup.backup.log (END) --------------------
2018-10-16T10:19:03.000666Z WSREP_SST: [ERROR] ****************************************************
2018-10-16T10:19:03.001333Z WSREP_SST: [ERROR] Cleanup after exit with status:22
2018-10-16T10:19:03.002588Z WSREP_SST: [DEBUG] Cleaning up temporary directories
```

Donor log

#2

Scenario: Catching up cluster (SST, IST)

```
2018-10-16T10:19:02.998914Z WSREP_SST: [ERROR] ********************** FATAL ERROR **********************
2018-10-16T10:19:02.999567Z WSREP_SST: [ERROR] xtrabackup finished with error: 1.  Check /opt/projects/percona/
dn1//innobackup.backup.log
---------- innobackup.backup.log (START) ------------------
xtrabackup: recognized server arguments: --datadir=/opt/projects/percona/merge/57-merge/installed/pxc-node/dn1 --log_bin=mysql-bin --server-id=2 --defaul
ts_group=mysqld --parallel=4
xtrabackup: recognized client arguments: --datadir=/opt/projects/percona/merge/57-merge/installed/pxc-node/dn1 --log_bin=mysql-bin --server-id=2 --defaul
ts_group=mysqld --parallel=4 --socket=/tmp/n1.sock --lock-ddl=1 --backup=1 --galera-info=1 --binlog-info=ON --stream=xbstream --target-dir=/tmp/pxc_sst_h
5zR/donor_xb_SE1L
encryption: using gcrypt 1.8.1
181016 15:49:02  version_check Connecting to MySQL server with DSN 'dbi:mysql:;mysql_read_default_group=xtrabackup;mysql_socket=/tmp/n1.sock' (using pass
word: NO).
Failed to connect to MySQL server as DBD::mysql module is not installed at - line 1327.
181016 15:49:02 Connecting to MySQL server ho
Failed to connect to MySQL server: Access den
---------- innobackup.backup.log (END) --
2018-10-16T10:19:03.000666Z WSREP_SST
2018-10-16T10:19:03.001333Z WSREP_SST
2018-10-16T10:19:03.002588Z WSREP_SST: [DEBUG] Cleaning up temporary directories
```

Donor log

Possible cause:

- Specified wsrep_sst_auth user doesn't exit.
- Credentials are wrong.
- Insufficient privileges.

#2

Scenario: Catching up cluster (SST, IST)

```
WSREP_SST: [ERROR] FATAL: PXC is receiving an SST from a node with a higher version. (2018-10-16 16:06:11)
WSREP_SST: [ERROR] This node's PXC version is 5.6.  The donor's PXC version is 5.7. (2018-10-16 16:06:11)
WSREP_SST: [ERROR] Upgrade this node before joining the cluster. (2018-10-16 16:06:11)
WSREP_SST: [ERROR] Cleanup after exit with status:2 (2018-10-16 16:06:11)
2018-10-16 16:06:11 17586 [ERROR] WSREP: Process completed with error: wsrep_sst_xtrabackup-v2 --role 'joiner' --address '127.0.0.1:5000' --datadir '/o
/projects/percona/merge/56-merge/installed/pxc-node/dn2/' --defaults-file '/opt/projects/percona/merge/56-merge/installed/pxc-node/n2.cnf' --defaults-g
up-suffix '' --parent '17586' --mysqld-version '5.6.41-84.1'  --binlog 'mysql-bin' : 2 (No such file or directory)
2018-10-16 16:06:11 17586 [ERROR] WSREP: Failed to read uuid:seqno from joiner script.
2018-10-16 16:06:11 17586 [ERROR] WSREP: SST script aborted with error 2 (No such file or directory)
2018-10-16 16:06:11 17586 [ERROR] WSREP: SST failed: 2 (No such file or directory)
2018-10-16 16:06:11 17586 [ERROR] Aborting
```

Joiner log

#3

Scenario: Catching up cluster (SST, IST)

```
WSREP_SST: [ERROR] FATAL: PXC is receiving an SST from a node with a higher version. (2018-10-16 16:06:11)
WSREP_SST: [ERROR] This node's PXC version is 5.6.  The donor's PXC version is 5.7. (2018-10-16 16:06:11)
WSREP_SST: [ERROR] Upgrade this node before joining the cluster. (2018-10-16 16:06:11)
WSREP_SST: [ERROR] Cleanup after exit with status:2 (2018-10-16 16:06:11)
2018-10-16 16:06:11 17586 [ERROR] WSREP: Process completed with error: wsrep_sst_xtrabackup-v2 --role 'joiner' --address '127.0.0.1:5000' --datadir '/o
/projects/percona/merge/56-merge/installed/pxc-node/dn2/' --defaults-file '/opt/projects/percona/merge/56-merge/installed/pxc-node/n2.cnf' --defaults-g
up-suffix '' --parent '17586' --mysqld-version '5.6.41-84.1'  --binlog 'mysql-bin' : 2 (No such file or directory)
2018-10-16 16:06:11 17586 [ERROR] WSREP: Failed to read uuid:seqno from joiner script.
2018-10-16 16:06:11 17586 [ERROR] WSREP: SST script aborted with error 2 (No such file or directory)
2018-10-16 16:06:11 17586 [ER
2018-10-16 16:06:11 17586 [ER
```

Joiner log

#3

Trying to get old version JOINER to join from new version DONOR. (Not supported).

Opposite is naturally allowed.

Scenario: Catching up cluster (SST, IST)

Donor log

```
2018/10/16 10:41:56 socat[8698] E connect(6, AF=2 10.0.2.15:4444, 16): Connection refused
        2018-10-16T10:41:56.957956Z WSREP_SST: [ERROR] ****************** FATAL ERROR **********************
        2018-10-16T10:41:56.962352Z WSREP_SST: [ERROR] Error while sending data to joiner node:  exit codes: 0 1
        2018-10-16T10:41:56.966502Z WSREP_SST: [ERROR] ****************************************************
        2018-10-16T10:41:56.970949Z WSREP_SST: [ERROR] Cleanup after exit with status:32
2018-10-16T10:41:56.982152Z 0 [ERROR] WSREP: Process completed with error: wsrep_sst_xtrabackup-v2 --role 'donor' --address '10.0.2.15:4444/xtrabackup_sst//1' --socket '/var/lib/mysql/mysql.sock' --datadir '/var/lib/mysql/' --defaults-file '/etc/my.cnf' --defaults-group-suffix '' --mysqld-version '5.7.23-23-57'   '' --gtid 'cb988f50-d12f-11e8-80c3-d781cb1f364b:0' : 32 (Broken pipe)
```

Joiner log

```
erminated
        2018-10-16T10:42:06.719766Z WSREP_SST: [ERROR] Removing /var/lib/mysql//xtrabackup_galera_info file due to signal
        2018-10-16T10:42:06.721798Z WSREP_SST: [ERROR] Removing  file due to signal
        2018-10-16T10:42:06.723709Z WSREP_SST: [ERROR] ****************** FATAL ERROR **********************
        2018-10-16T10:42:06.724708Z WSREP_SST: [ERROR] Error while getting data from donor node:  exit codes: 143 143
        2018-10-16T10:42:06.725747Z WSREP_SST: [ERROR] ****************************************************
        2018-10-16T10:42:06.726930Z WSREP_SST: [ERROR] Cleanup after exit with status:32
```

#4

Scenario: Catching up cluster (SST, IST)

Donor log

```
2018/10/16 10:41:56 socat[8698] E connect(6, AF=2 10.0.2.15:4444, 16): Connection refused
        2018-10-16T10:41:56.957956Z WSREP_SST: [ERROR] ****************** FATAL ERROR **********************
        2018-10-16T10:41:56.962352Z WSREP_SST: [ERROR] Error while sending data to joiner node:  exit codes: 0 1
        2018-10-16T10:41:56.966502Z WSREP_SST: [ERROR] ****************************************************
        2018-10-16T10:41:56.970949Z WSREP_SST: [ERROR] Cleanup after exit with status:32
2018-10-16T10:41:56.982152Z 0 [ERROR] WSREP: Process completed with error: wsrep_sst_xtrabackup-v2 --role 'donor' --address '10.0.2.15:4444/xtrabackup_sst//1' --socket '/var/lib/mysql/mysql.sock' --datadir '/var/lib/mysql/' --defaults-file '/etc/my.cnf' --defaults-group-suffix '' --mysqld-version '5.7.23-23-57'   '' --gtid 'cb988f50-d12f-11e8-80c3-d781cb1f364b:0' : 32 (Broken pipe)
```

Joiner log

WSREP_SST: [WARNING] wsrep_node_address or wsrep_sst_receive_address not set. Consider setting them if SST fails.

```
2018-10-16T10:42:06.726930Z WSREP_SST: [ERROR] Cleanup after exit with status:32
```

#4

Scenario: Catching up cluster (SST, IST)

```
2018-10-16T10:50:13.294596Z WSREP_SST: [DEBUG] The xtrabackup version is 2.4.12
2018-10-16T10:50:13.339114Z WSREP_SST: [WARNING] wsrep_node_address or wsrep_sst_receive_address not set. Consider setting them if SST fails.
2018-10-16T10:50:13.445950Z WSREP_SST: [DEBUG] Streaming with xbstream
2018-10-16T10:50:13.446827Z WSREP_SST: [DEBUG] Using socat as streamer
2018-10-16T10:50:13.453913Z WSREP_SST: [DEBUG] Using openssl based encryption with socat: with key, crt, and ca
2018-10-16T10:50:13.466450Z WSREP_SST: [DEBUG] Encrypting with CERT: server-cert.pem, KEY: server-key.pem, CA: ca.pem
2018-10-16T10:50:13.486102Z WSREP_SST: [DEBUG] Streaming SST meta-info file before SST
2018-10-16T10:50:13.487131Z WSREP_SST: [DEBUG] Evaluating (@ donor-OpenSSL-Encrypted-4-sst-info) xbstream $xbstreameopts -c ${FILE_TO_STREAM} | s
cat -u stdio openssl-connect:127.0.0.1:5000,cert=server-cert.pem,key=server-key.pem,cafile=ca.pem,verify=1,commonname="",retry=30; RC=( ${PIPESTATUS[@]}
)
ncryption: using gcrypt 1.8.1
018/10/16 16:20:13 socat[21333] E SSL_CTX_load_verify_locations(): error:02001002:system library:fopen:No such file or directory
2018-10-16T10:50:13.491298Z WSREP_SST: [ERROR] ****************** FATAL ERROR *********************
2018-10-16T10:50:13.492319Z WSREP_SST: [ERROR] Error while sending data to joiner node:  exit codes: 0 1
2018-10-16T10:50:13.493317Z WSREP_SST: [ERROR] ****************************************************
2018-10-16T10:50:13.494472Z WSREP_SST: [ERROR] Cleanup after exit with status:32
2018-10-16T10:50:13.496395Z WSREP_SST: [DEBUG] Cleaning up temporary directories
```

#5

Scenario: Catching up cluster (SST, IST)

```
2018-10-16T10:50:13.294596Z WSREP_SST: [DEBUG] The xtrabackup version is 2.4.12
2018-10-16T10:50:13.339114Z WSREP_SST: [WARNING] wsrep_node_address or wsrep_sst_receive_address not set. Consider setting them if SST fails.
2018-10-16T10:50:13.445950Z WSREP_SST: [DEBUG] Streaming with xbstream
2018-10-16T10:50:13.446827Z WSREP_SST: [DEBUG] Using socat as streamer
2018-10-16T10:50:13.453913Z WSREP_SST: [DEBUG] Using openssl based encryption with socat: with key, crt, and ca
2018-10-16T10:50:13.466450Z WSREP_SST: [DEBUG] Encrypting with CERT: server-cert.pem, KEY: server-key.pem, CA: ca.pem
2018-10-16T10:50:13.486102Z WSREP_SST: [DEBUG] Streaming SST meta-info file before SST
2018-10-16T10:50:13.487131Z WSREP_SST: [DEBUG] Evaluating (@ donor-OpenSSL-Encrypted-4-sst-info) xbstream $xbstreameopts -c ${FILE_TO_STREAM} | s
cat -u stdio openssl-connect:127.0.0.1:5000,cert=server-cert.pem,key=server-key.pem,cafile=ca.pem,verify=1,commonname="",retry=30; RC=( ${PIPESTATUS[@]}
)
ncryption: using gcrypt 1.8.1
018/10/16 16:20:13 socat[21333] E SSL_CTX_load_verify_locations(): error:02001002:system library:fopen:No such file or directory
2018-10-16T10:50:13.491298Z WSREP_SST: [ERROR] ********** FATAL ERROR **********
2018-10-16T10:50:13.492319Z WSREP_SST: [ERROR]
2018-10-16T10:50:13.493317Z WSREP_SST: [ERROR]
2018-10-16T10:50:13.494472Z WSREP_SST: [ERROR]
2018-10-16T10:50:13.496395Z WSREP_SST: [DEBUG]
```

Faulty SSL configuration

#5

Scenario: Catching up cluster (SST, IST)

PXC recommends: **Same configuration on all nodes of the cluster.**

Old DONOR - New JOINER (OK)

XB is external tool and has its own set of controllable configuration (passed through PXC my.cnf)

SST user should be present on DONOR

Look at DONOR and JOINER log.

wsrep_sst_recieve_address/wsrep_node_address is needed.

Advance encryption option like keyring on DONOR and no keyring on JOINER is not allowed.

Ensure stable n/w link between DONOR and JOINER.

Network rules (firewall, etc..). SST uses port 4444. IST uses 4568.

Often-error are local to XB. Check the XB log file that can give hint of error.

Scenario: Cluster doesn’t come up on restart

Scenario: Cluster doesn't come up on restart

- All your nodes are located in same Data-Center (DC)
- DC hits power failure and all nodes are restarted.
- On restart, recovery flow is executed to recover wsrep coordinates.

Scenario: Cluster doesn't come up on restart

- All your nodes are located in same Data-Center (DC)
- DC hits power failure and all nodes are restarted.
- On restart, recovery flow is executed to recover wsrep coordinates.

```
018-10-15T03:02:09.625988Z 0 [Note] WSREP: Before binlog recovery (wsrep position cc20310c-cdee-11e8-b1bb-67cd8f61d62d:7)
018-10-15T03:02:09.626003Z 0 [Note] Starting crash recovery...
018-10-15T03:02:09.626087Z 0 [Note] Crash recovery finished.
018-10-15T03:02:09.626096Z 0 [Note] WSREP: After binlog recovery (wsrep position cc20310c-cdee-11e8-b1bb-67cd8f61d62d:7)
018-10-15T03:02:09.626254Z 0 [Note] WSREP: Recovered position: cc20310c-cdee-11e8-b1bb-67cd8f61d62d:7
018-10-15T03:02:09.626263Z 0 [Note] Binlog end
018-10-15T03:02:09.626299Z 0 [Note] Shutting down plugin 'ngram'
018-10-15T03:02:09.626305Z 0 [Note] Shutting down plugin 'partition'
```

Scenario: Cluster doesn't come up on restart

- All your nodes are located in same Data-Center (DC)
- DC hits power failure and all nodes are restarted.
- On restart, recovery flow is executed to recover wsrep coordinates.

```
018-10-15T03:02:09.625988Z 0 [Note] WSREP: Before binlog recovery (wsrep position cc20310c-cdee-11e8-b1bb-67cd8f61d62d:7)
018-10-15T03:02:09.626003Z 0 [Note] Starting crash recovery...
018-10-15T03:02:09.626087Z 0 [Note] Crash recovery finished.
018-10-15T03:02:09.626096Z 0 [Note] WSREP: After binlog recovery (wsrep position cc20310c-cdee-11e8-b1bb-67cd8f61d62d:7)
018-10-15T03:02:09.626254Z 0 [Note] WSREP: Recovered position: cc20310c-cdee-11e8-b1bb-67cd8f61d62d:7
018-10-15T03:02:09.626263Z 0 [Note] Binlog end
018-10-15T03:02:09.626299Z 0 [Note] Shutting down plugin 'ngram'
018-10-15T03:02:09.626305Z 0 [Note] Shutting down plugin 'partition'
```

Cluster still fails to come up

Scenario: Cluster doesn't come up on restart

- Close look at the log shows original bootstrapping node has `safe_to_bootstrap` set to 0 so it refuse to come up.

```
[root@pxc1 mysql]# cat grastate.dat
# GALERA saved state
version: 2.1
uuid:    cc20310c-cdee-11e8-b1bb-67cd8f61d62d
seqno:   7
safe_to_bootstrap: 0
  [root@pxc1 mysql]#
```

- Other nodes of cluster are left dangling (in non-primary state) in absence of original cluster forming node.

Scenario: Cluster doesn't come up on restart

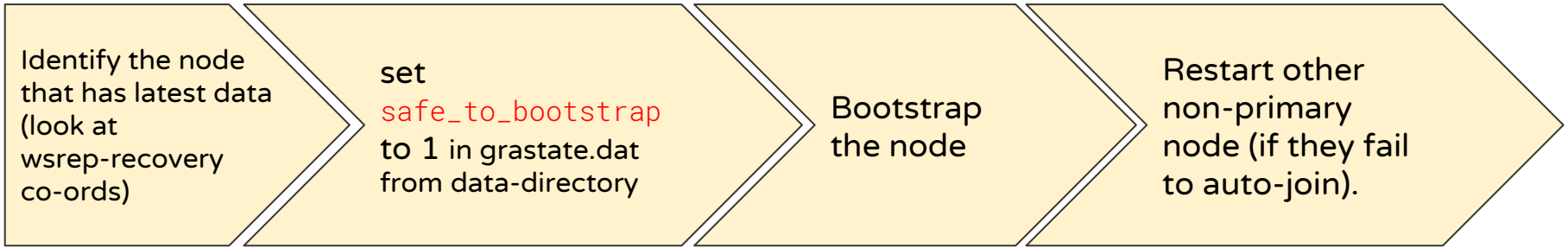
- Close look at the log shows original bootstrapping node has `safe_to_bootstrap` set to 0 so it refuse to come up.

```
[root@pxc1 mysql]# cat grastate.dat
# GALERA saved state
version: 2.1
uuid:    cc20310c-cdee-11e8-b1bb-67cd8f61d62d
seqno:   7
safe_to_bootstrap: 0
  [root@pxc1 mysql]#
```

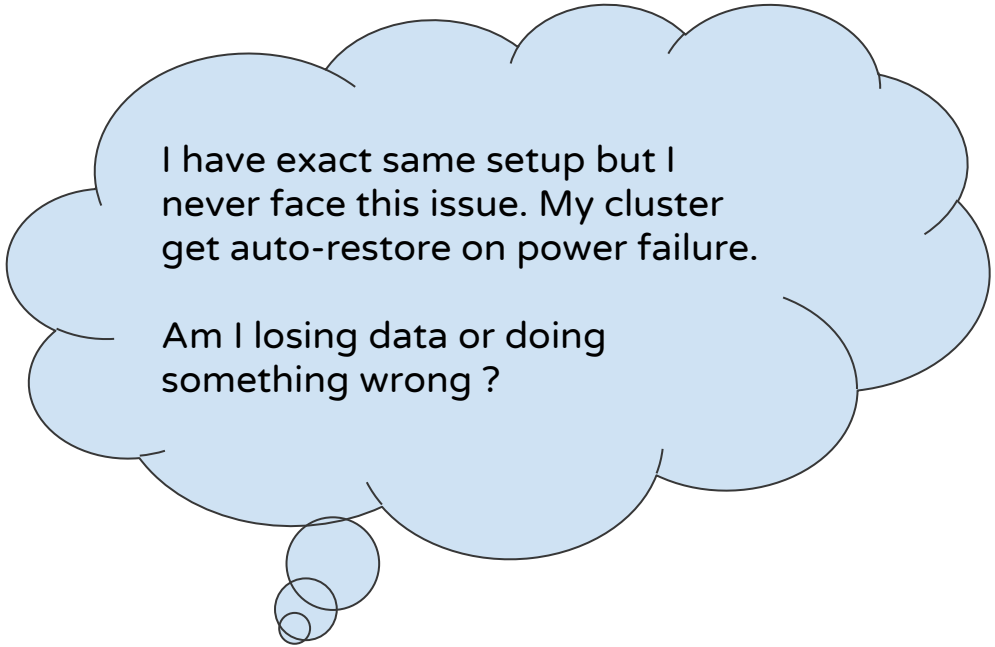
- Other nodes of cluster are lef
  absence of original cluster form

Galera/PXC expect user to identify node that has latest data and then use that too bootstrap. So as safety check `safe_to_bootstrap` was added.

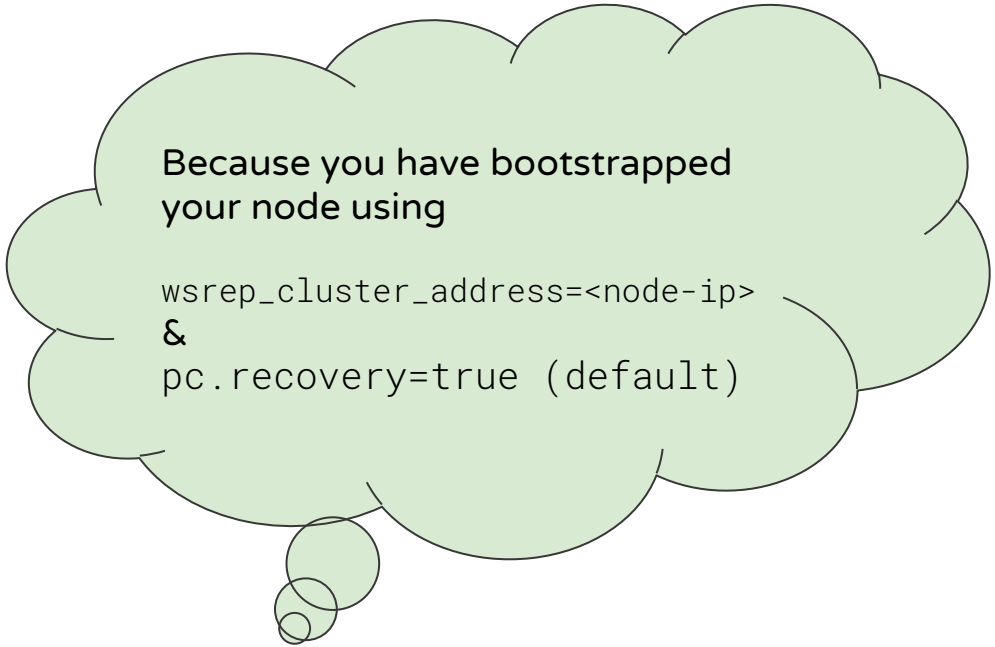
Scenario: Cluster doesn't come up on restart



Scenario: Cluster doesn't come up on restart



Scenario: Cluster doesn't come up on restart



Scenario: Cluster doesn't come up on restart

Error is observed if you have bootstrapped:

`wsrep_cluster_address="gcomm://"`

OR
`wsrep_cluster_address="<node-ips>"`

but `pc.recovery=false`

Because you have bootstrapped your node using

`wsrep_cluster_address=<node-ip>`
&
`pc.recovery=true (default)`

Scenario: Cluster doesn't come up on restart

PXC can auto-restart on DC failure depending on configuration option used.

Scenario: Data inconsistency

Scenario: Data inconsistency

```
2018-10-15T09:04:59.368182Z 7 [ERROR] Slave SQL: Could not execute Update_rows event on table test.sbtest1; Can't find record in 'sbtest1', Error_code: 1
032; handler error HA_ERR_KEY_NOT_FOUND; the event's master log FIRST, end_log_pos 536, Error_code: 1032
2018-10-15T09:04:59.368214Z 7 [Warning] WSREP: RBR event 3 Update_rows apply warning: 120, 255823
2018-10-15T09:04:59.368361Z 7 [Note] WSREP: Applier statement rollback needed
2018-10-15T09:04:59.368391Z 7 [Warning] WSREP: Failed to apply app buffer: seqno: 255823, status: 1
         at galera/src/trx_handle.cpp:apply():353
Retrying 2th time
2018-10-15T09:04:59.368453Z 7 [ERROR] Slave SQL: Could not execute Update_rows event on table test.sbtest1; Can't find record in 'sbtest1', Error_code: 1
032; handler error HA_ERR_KEY_NOT_FOUND; the event's master log FIRST, end_log_pos 536, Error_code: 1032
2018-10-15T09:04:59.368468Z 7 [Warning] WSREP: RBR event 3 Update_rows apply warning: 120, 255823
2018-10-15T09:04:59.368694Z 7 [Note] WSREP: Applier statement rollback needed
2018-10-15T09:04:59.368726Z 7 [Warning] WSREP: Failed to apply app buffer: seqno: 255823, status: 1
         at galera/src/trx_handle.cpp:apply():353
Retrying 3th time
2018-10-15T09:04:59.368784Z 7 [ERROR] Slave SQL: Could not execute Update_rows event on table test.sbtest1; Can't find record in 'sbtest1', Error_code: 1
032; handler error HA_ERR_KEY_NOT_FOUND; the event's master log FIRST, end_log_pos 536, Error_code: 1032
2018-10-15T09:04:59.368798Z 7 [Warning] WSREP: RBR event 3 Update_rows apply warning: 120, 255823
2018-10-15T09:04:59.369342Z 7 [Note] WSREP: Applier statement rollback needed
2018-10-15T09:04:59.369384Z 7 [Warning] WSREP: Failed to apply app buffer: seqno: 255823, status: 1
         at galera/src/trx_handle.cpp:apply():353
Retrying 4th time
2018-10-15T09:04:59.369571Z 7 [ERROR] Slave SQL: Could not execute Update_rows event on table test.sbtest1; Can't find record in 'sbtest1', Error_code: 1
032; handler error HA_ERR_KEY_NOT_FOUND; the event's master log FIRST, end_log_pos 536, Error_code: 1032
2018-10-15T09:04:59.369587Z 7 [Warning] WSREP: RBR event 3 Update_rows apply warning: 120, 255823
2018-10-15T09:04:59.369960Z 7 [Note] WSREP: Applier statement rollback needed
2018-10-15T09:04:59.370670Z 7 [ERROR] WSREP: Failed to apply trx: source: 3bf87f07-d055-11e8-bcdd-daf9833d0a6c version: 4 local: 0 state: APPLYING flags:
 1 conn_id: 15 trx_id: 213109 seqnos (l: 72280, g: 255823, s: 255822, d: 255708, ts: 1792600291956)
2018-10-15T09:04:59.370693Z 7 [ERROR] WSREP: Failed to apply trx 255823 4 times
2018-10-15T09:04:59.370700Z 7 [ERROR] WSREP: Node consistency compromised, aborting...
2018-10-15T09:04:59.370732Z 7 [Note] WSREP: turning isolation on
2018-10-15T09:04:59.370818Z 7 [Note] WSREP: Closing send monitor...
2018-10-15T09:04:59.370826Z 7 [Note] WSREP: Closed send monitor.
2018-10-15T09:04:59.370833Z 7 [Note] WSREP: gcomm: terminating thread
2018-10-15T09:04:59.370838Z 7 [Note] WSREP: gcomm: joining thread
```

Scenario: Data inconsistency

- 2 kinds of inconsistencies
  - Physical inconsistency: Hardware Issues
  - Logical inconsistency: Data Issues

Scenario: Data inconsistency

- 2 kinds of inconsistencies
  - Physical inconsistency: Hardware Issues
  - Logical inconsistency: Data Issues

Logical inconsistency caused to cluster specific operation like locks, RSU, wsrep_on=off, etc…

Scenario: Data inconsistency

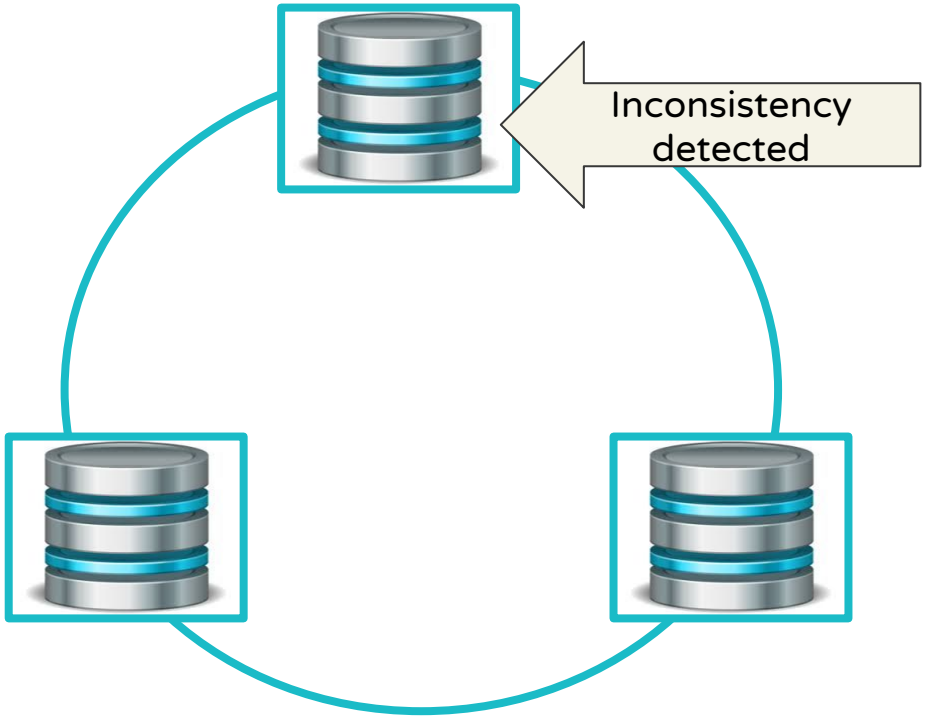
- 2 kinds of inconsistencies
  - Physical inconsistency: Hardware Issues
  - Logical inconsistency: Data Issues

Logical inconsistency caused to cluster specific operation like locks, RSU, wsrep_on=off, etc…

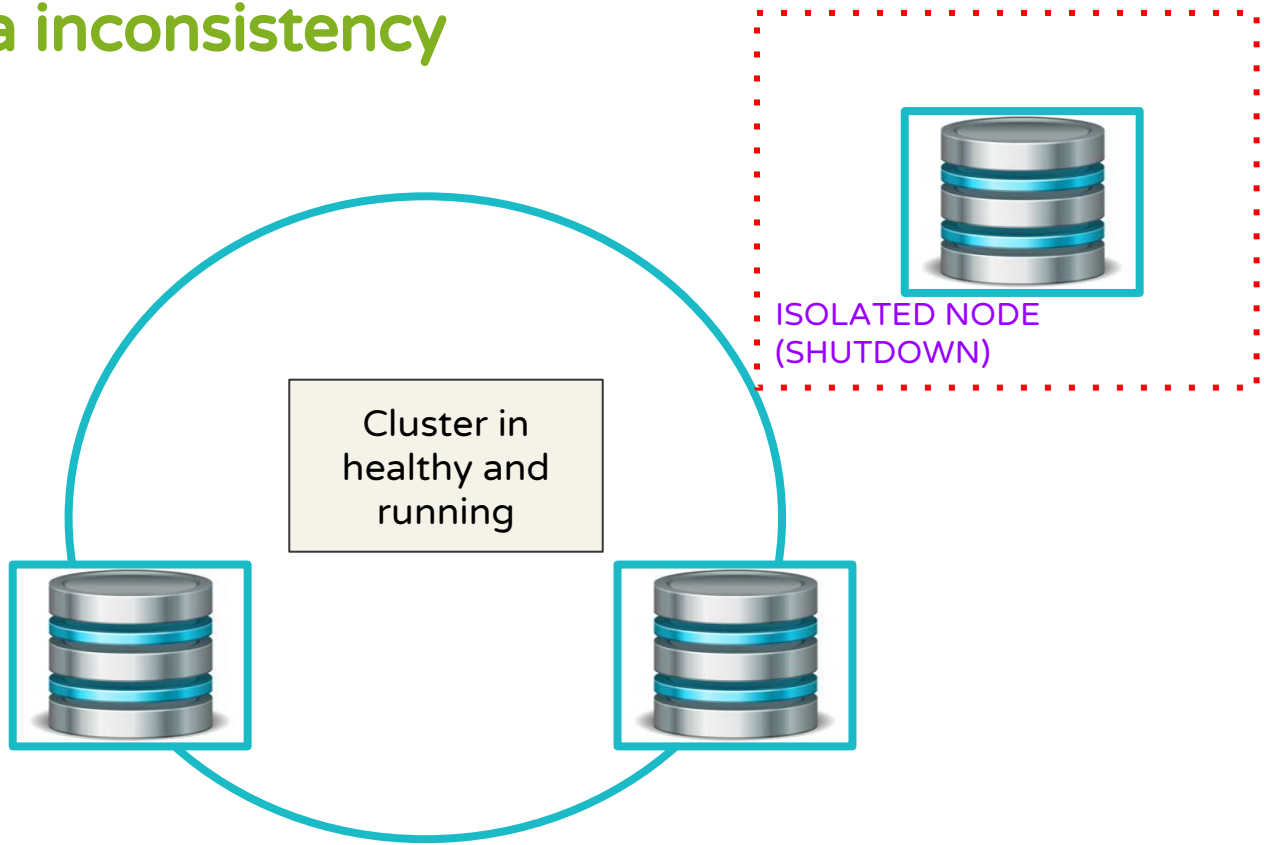
PXC has zero tolerance for inconsistency

and so it immediately isolate the nodes on detecting inconsistency.

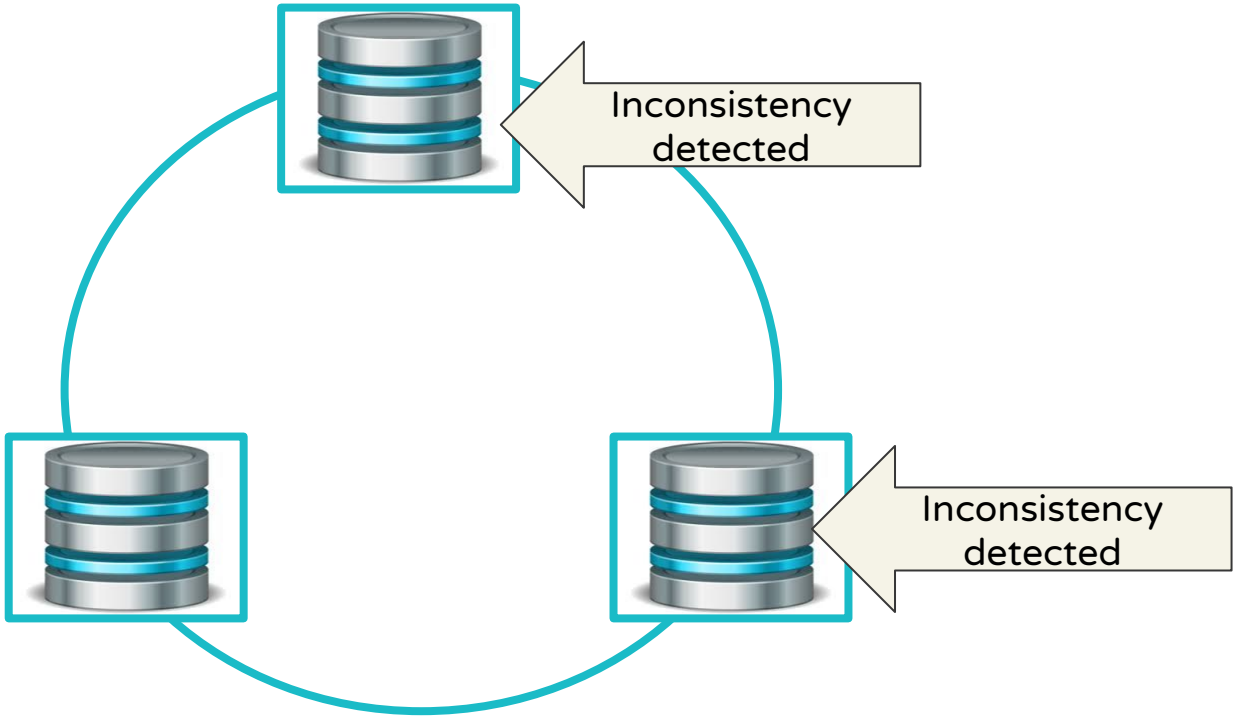
Scenario: Data inconsistency



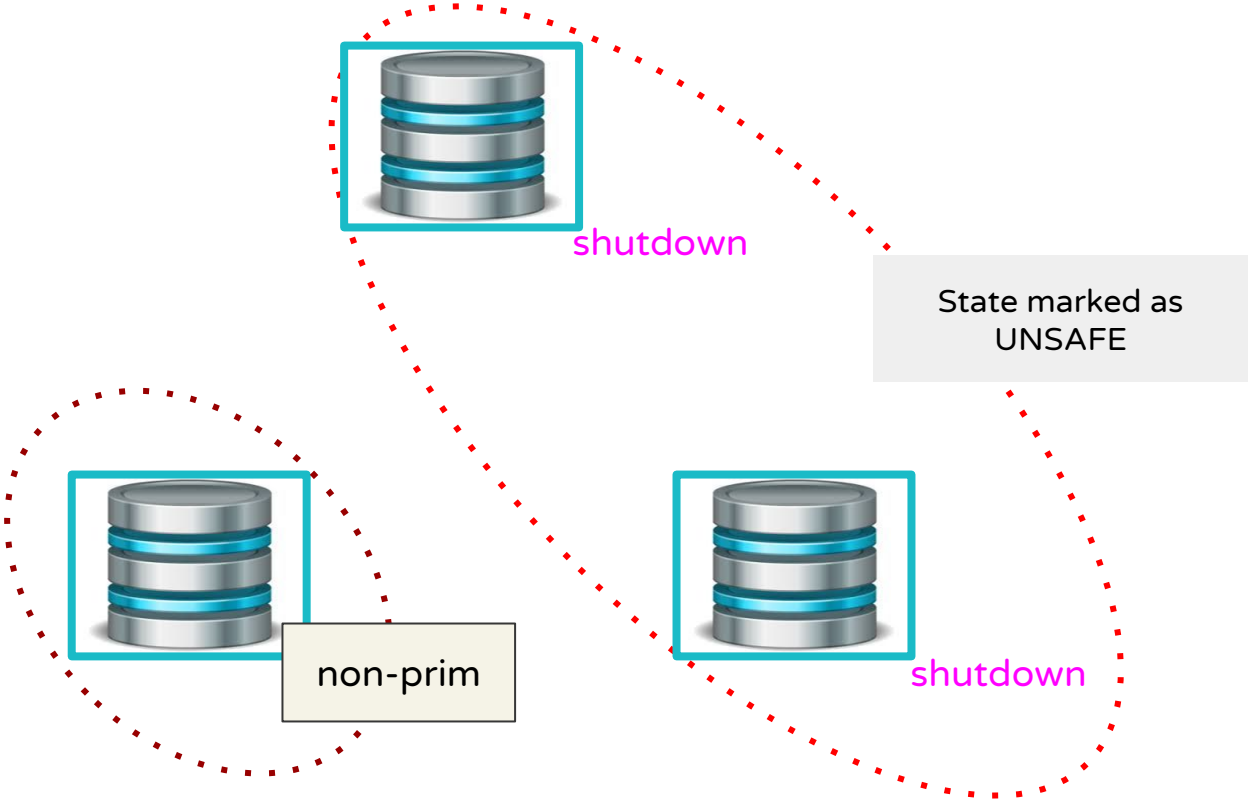
Scenario: Data inconsistency



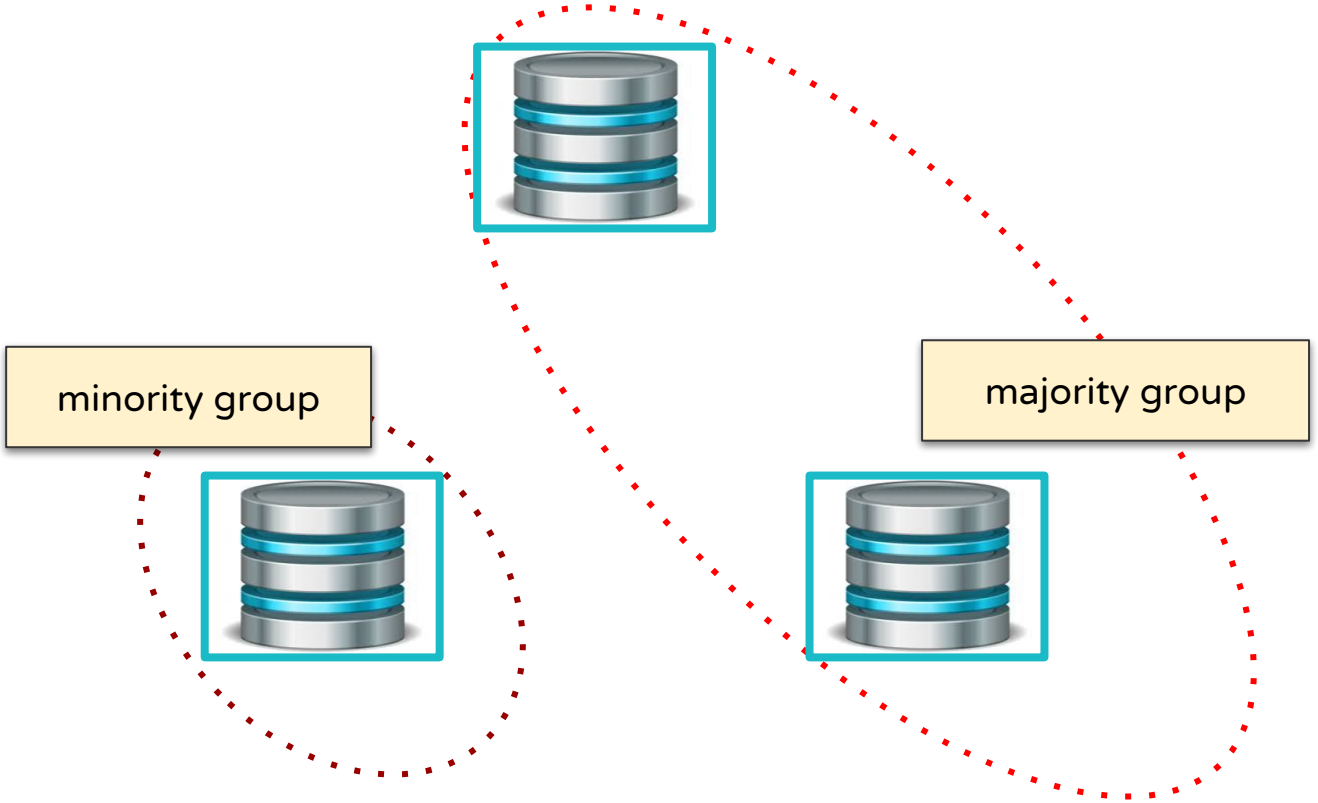
Scenario: Data inconsistency



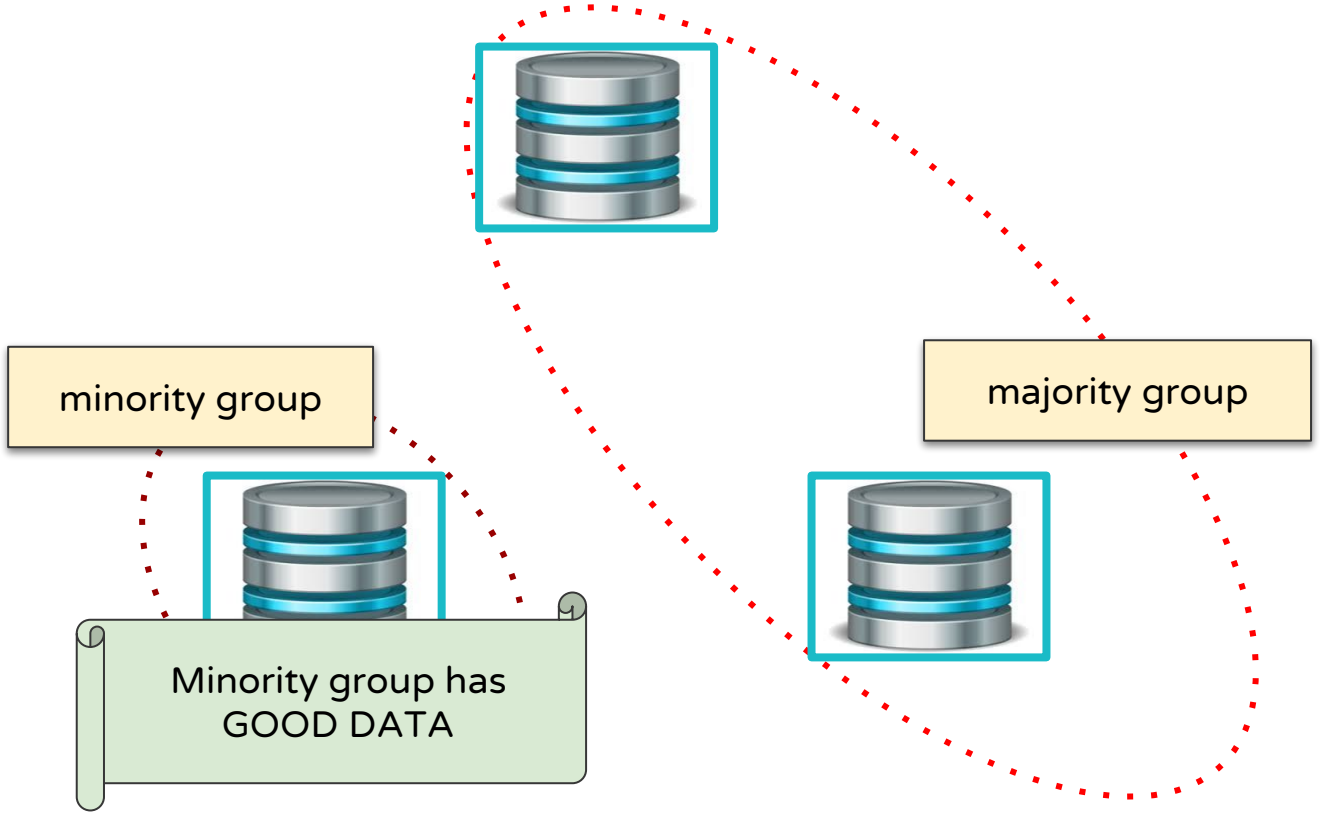
Scenario: Data inconsistency



Scenario: Data inconsistency

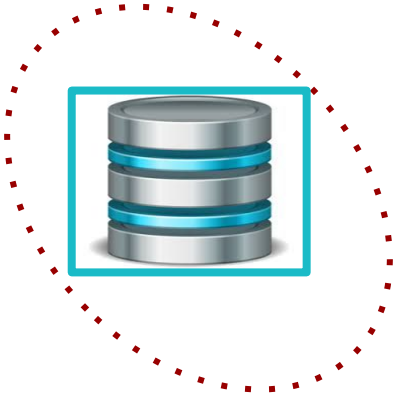


Scenario: Data inconsistency



Scenario: Data inconsistency

If there are multiple nodes in minority group, identify a node that has latest data.



Scenario: Data inconsistency

If there are multiple nodes in minority group, identify a node that has latest data.

Set `pc.bootstrap=1` on the selected node.

Single node cluster formed

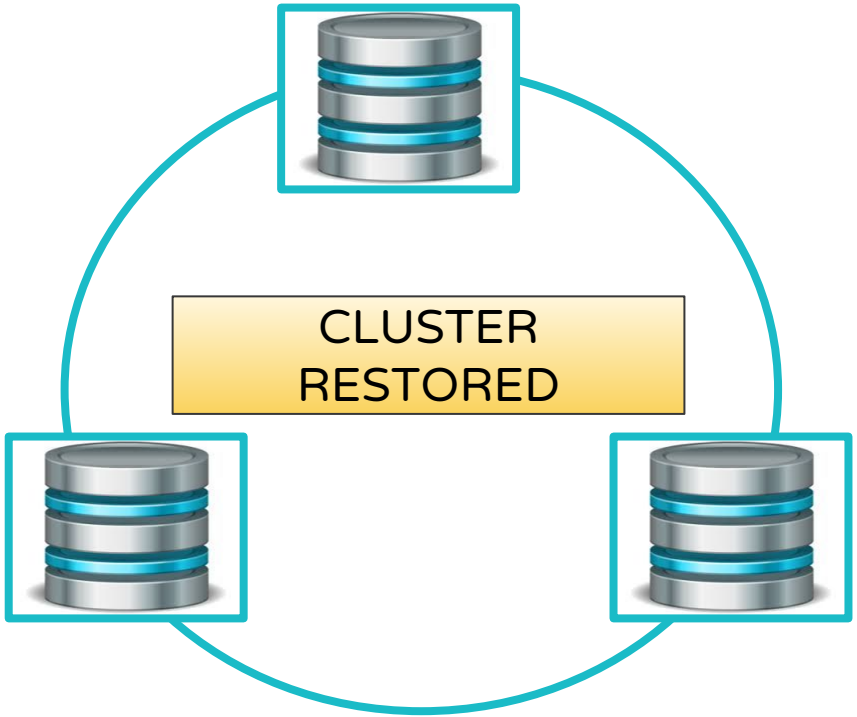
Scenario: Data inconsistency

If there are multiple nodes in minority group, identify a node that has latest data.

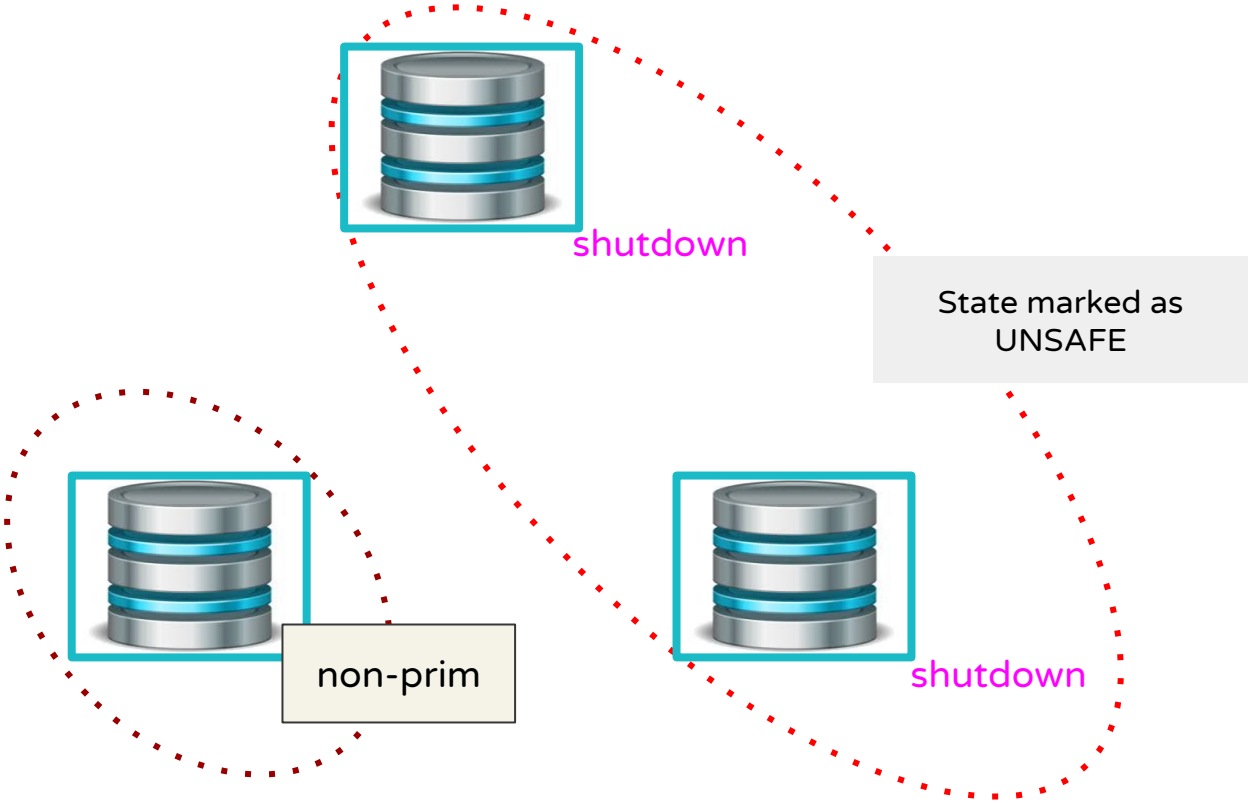
Set `pc.bootstrap=1` on the selected node.

Boot other majority node. (they will join through SST).

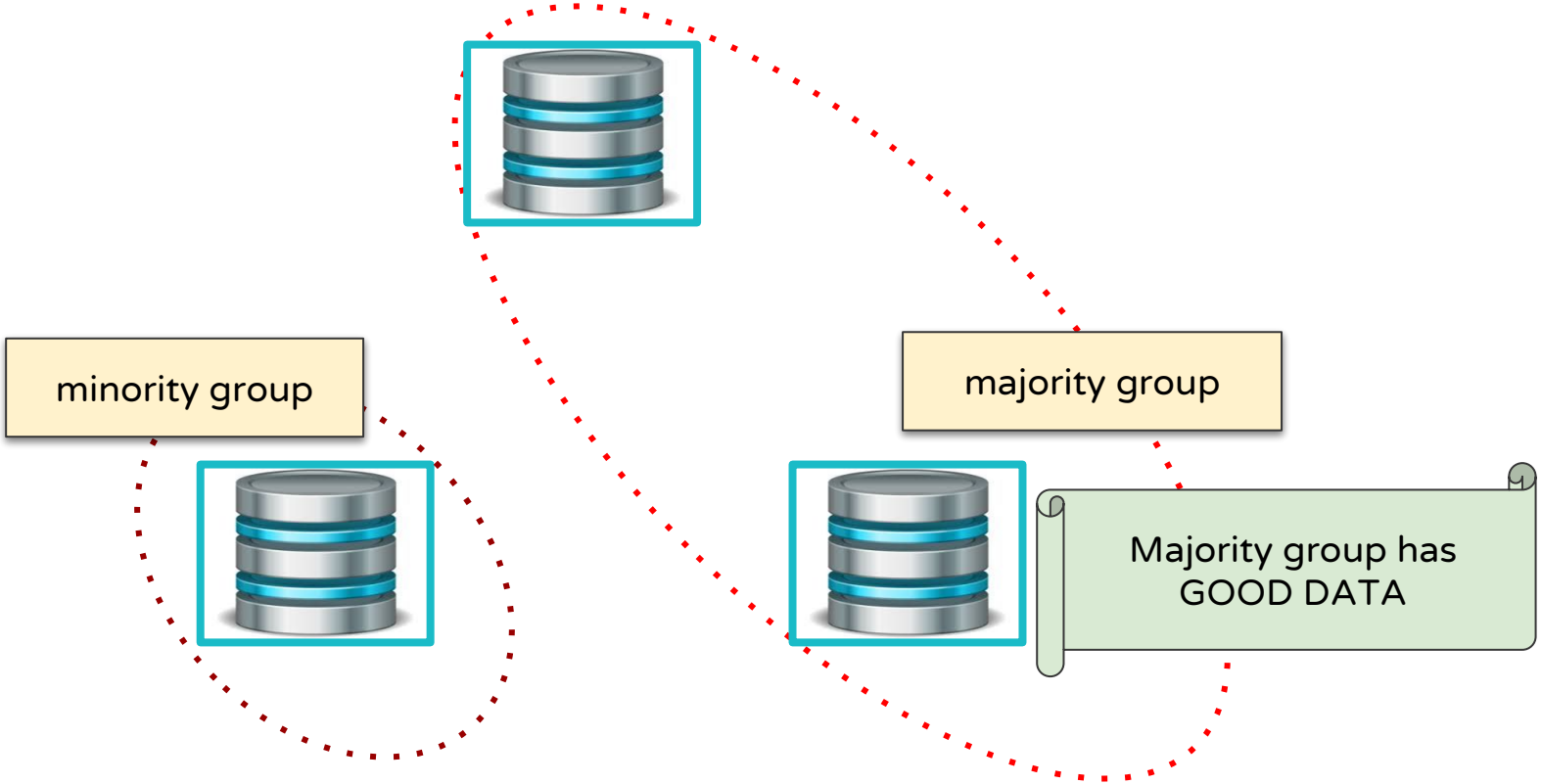
Scenario: Data inconsistency



Scenario: Data inconsistency

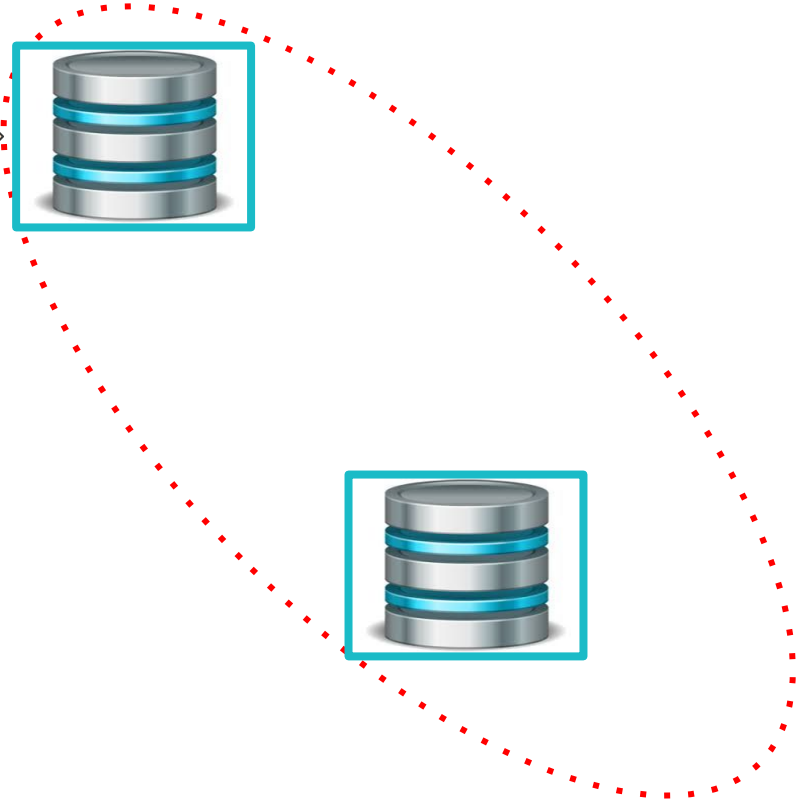


Scenario: Data inconsistency



Scenario: Data inconsistency

Nodes in majority group are already SHUTDOWN. Initiate SHUTDOWN of nodes from minority group.

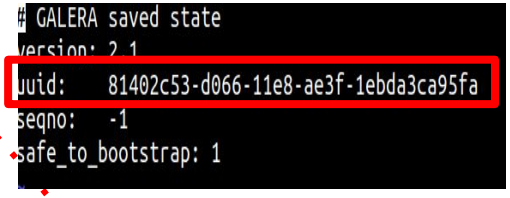


Scenario: Data inconsistency

Nodes in majority group are already SHUTDOWN. Initiate SHUTDOWN of nodes from minority group.

Fix grastate.dat for the nodes from majority group. (Consistency shutdown sequence has marked STATE=UNSAFE).

```
# GALERA saved state
version: 2.1
uuid:    00000000-0000-0000-0000-000000000000
seqno:   -1
safe_to_bootstrap: 0
```



```
# GALERA saved state
version: 2.1
uuid:    81402c53-d066-11e8-ae3f-1ebda3ca95fa
seqno:   -1
safe_to_bootstrap: 1
```

Valid uuid can be copied over from a minority group node.

Scenario: Data inconsistency

Nodes in majority group are already SHUTDOWN. Initiate SHUTDOWN of nodes from minority group.

Fix grastate.dat for the nodes from majority group. (Consistency shutdown sequence has marked STATE=UNSAFE).

Bootstrap the cluster using one of the node from majority group and eventually get other majority nodes to join.

Scenario: Data inconsistency

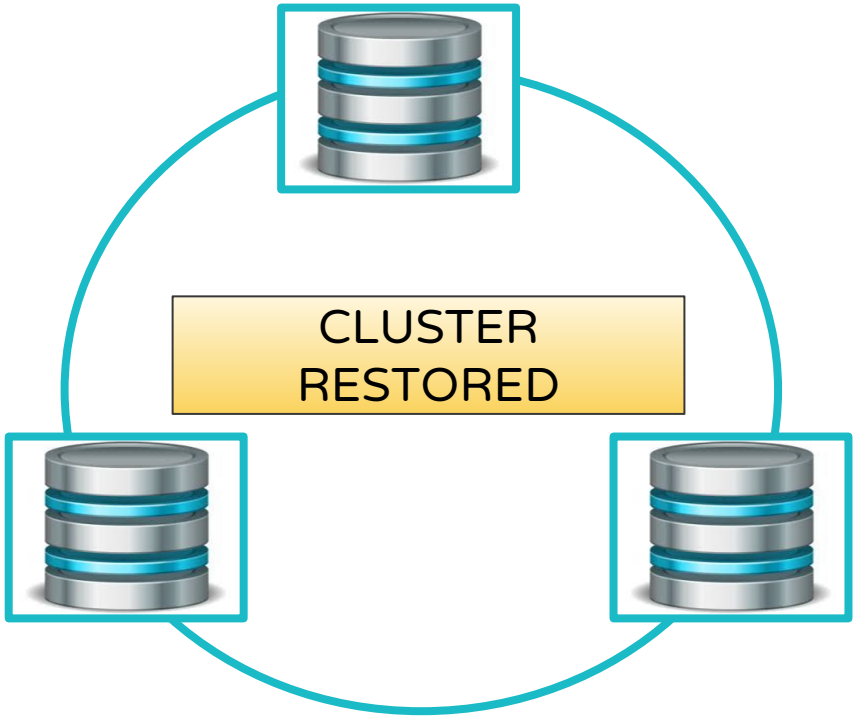
Nodes in majority group are already SHUTDOWN. Initiate SHUTDOWN of nodes from minority group.

Fix grastate.dat for the nodes from majority group. (Consistency shutdown sequence has marked STATE=UNSAFE).

Bootstrap the cluster using one of the node from majority group and eventually get other majority nodes to join.

Remove grastate.dat of minority group nodes and restart them to join newly formed cluster.

Scenario: Data inconsistency



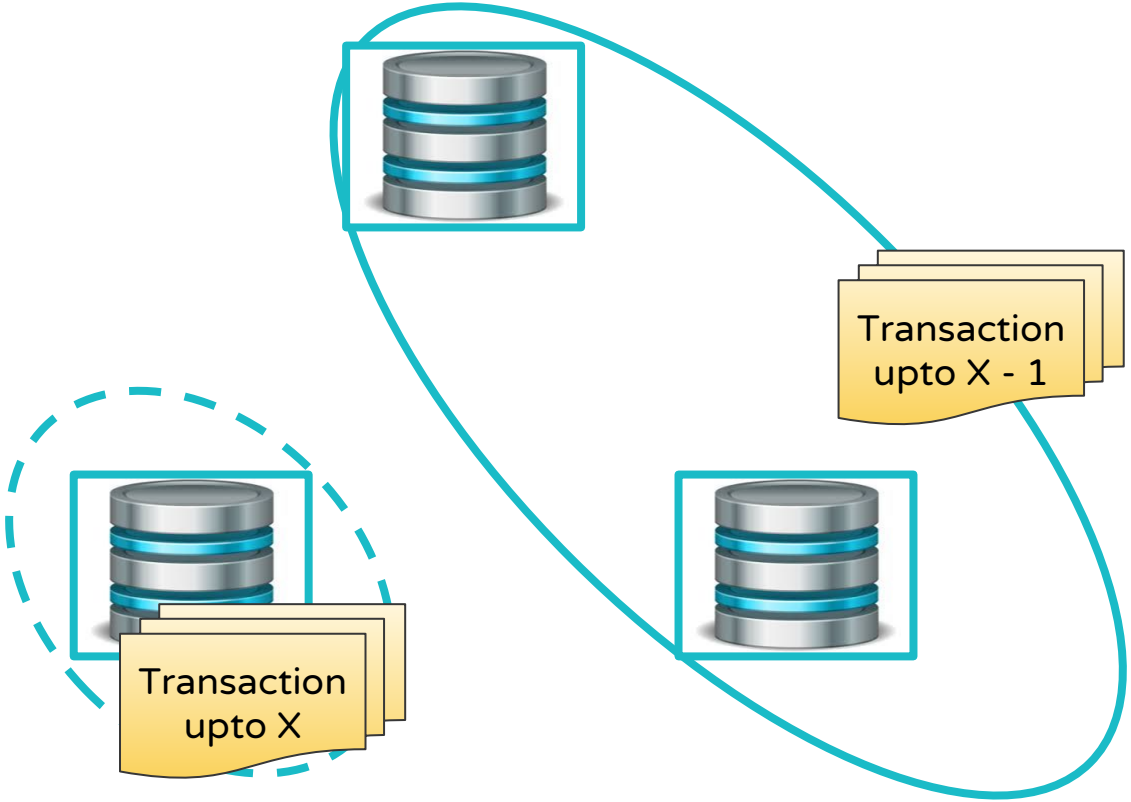
Scenario: Another aspect of data inconsistency

Scenario: Another aspect of data inconsistency

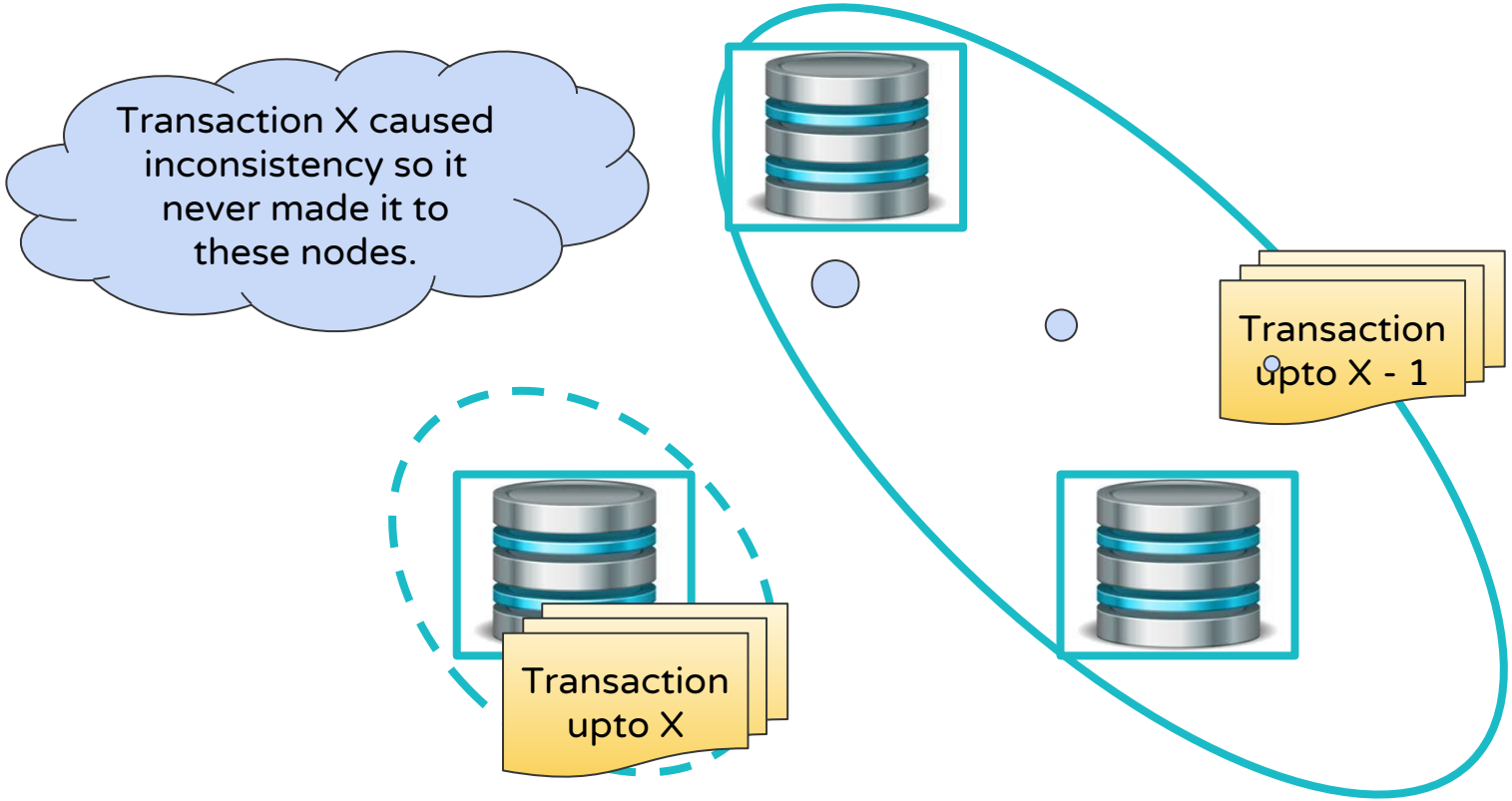
```
2018-10-15T10:46:41.798782Z 0 [Note] WSREP: Save the discovered primary-component to disk
2018-10-15T10:46:42.293055Z 0 [Note] WSREP: gcomm: connected
2018-10-15T10:46:42.293409Z 0 [Note] WSREP: Shifting CLOSED -> OPEN (TO: 0)
2018-10-15T10:46:42.293649Z 0 [Note] WSREP: Waiting for SST/IST to complete.
2018-10-15T10:46:42.293835Z 0 [Note] WSREP: New COMPONENT: primary = yes, bootstrap = no, my_idx = 2, memb_num = 3
2018-10-15T10:46:42.293876Z 0 [Note] WSREP: STATE EXCHANGE: Waiting for state UUID.
2018-10-15T10:46:42.294087Z 0 [Note] WSREP: STATE EXCHANGE: sent state msg: 99f54c0e-d067-11e8-b554-36e8108cdc10
2018-10-15T10:46:42.294121Z 0 [Note] WSREP: STATE EXCHANGE: got state msg: 99f54c0e-d067-11e8-b554-36e8108cdc10 from 0 (pxc-cluster-node-2)
2018-10-15T10:46:42.294140Z 0 [Note] WSREP: STATE EXCHANGE: got state msg: 99f54c0e-d067-11e8-b554-36e8108cdc10 from 1 (pxc-cluster-node-3)
2018-10-15T10:46:42.297310Z 0 [Note] WSREP: STATE EXCHANGE: got state msg: 99f54c0e-d067-11e8-b554-36e8108cdc10 from 2 (pxc-cluster-node-1)
2018-10-15T10:46:42.297386Z 0 [ERROR] WSREP: gcs/src/gcs_group.cpp:group_post_state_exchange():322: Reversing history: 2 -> 1, this member has applied 1
more events than the primary component.Data loss is possible. Aborting.
2018-10-15T10:46:42.297412Z 0 [Note] WSREP: /usr/sbin/mysqld: Terminated.
```

One of the node from minority group

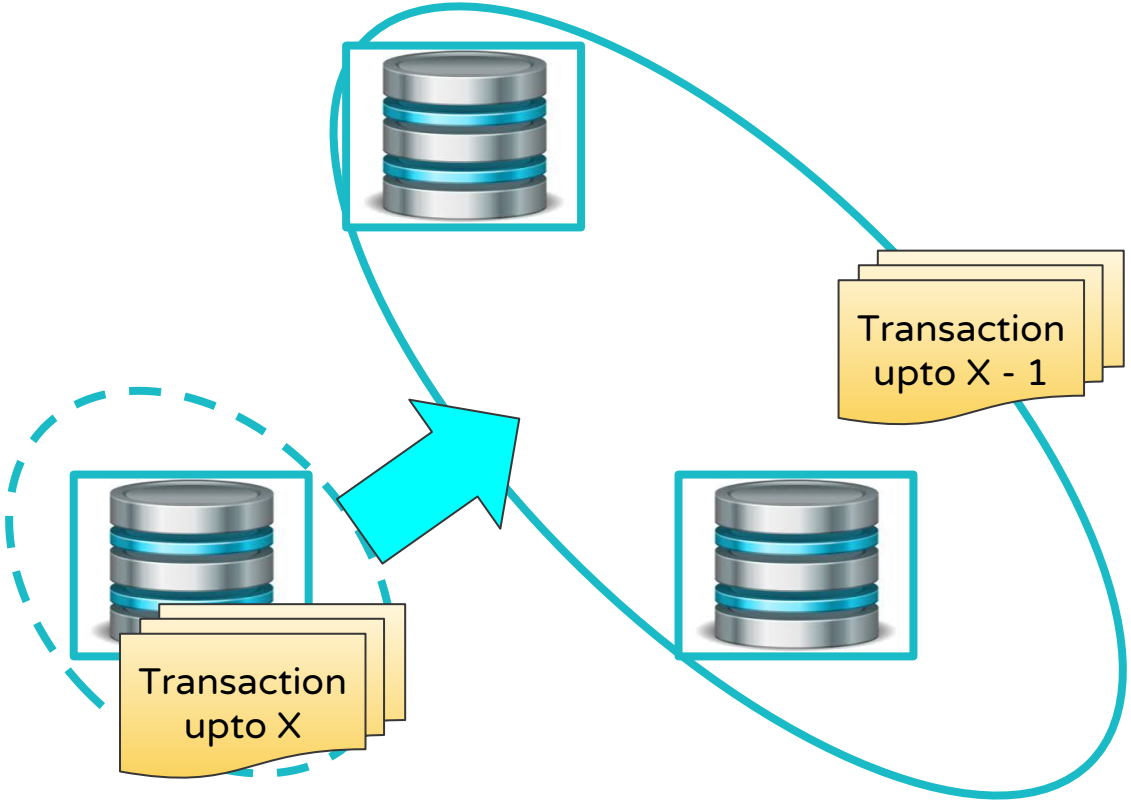
Scenario: Another aspect of data inconsistency



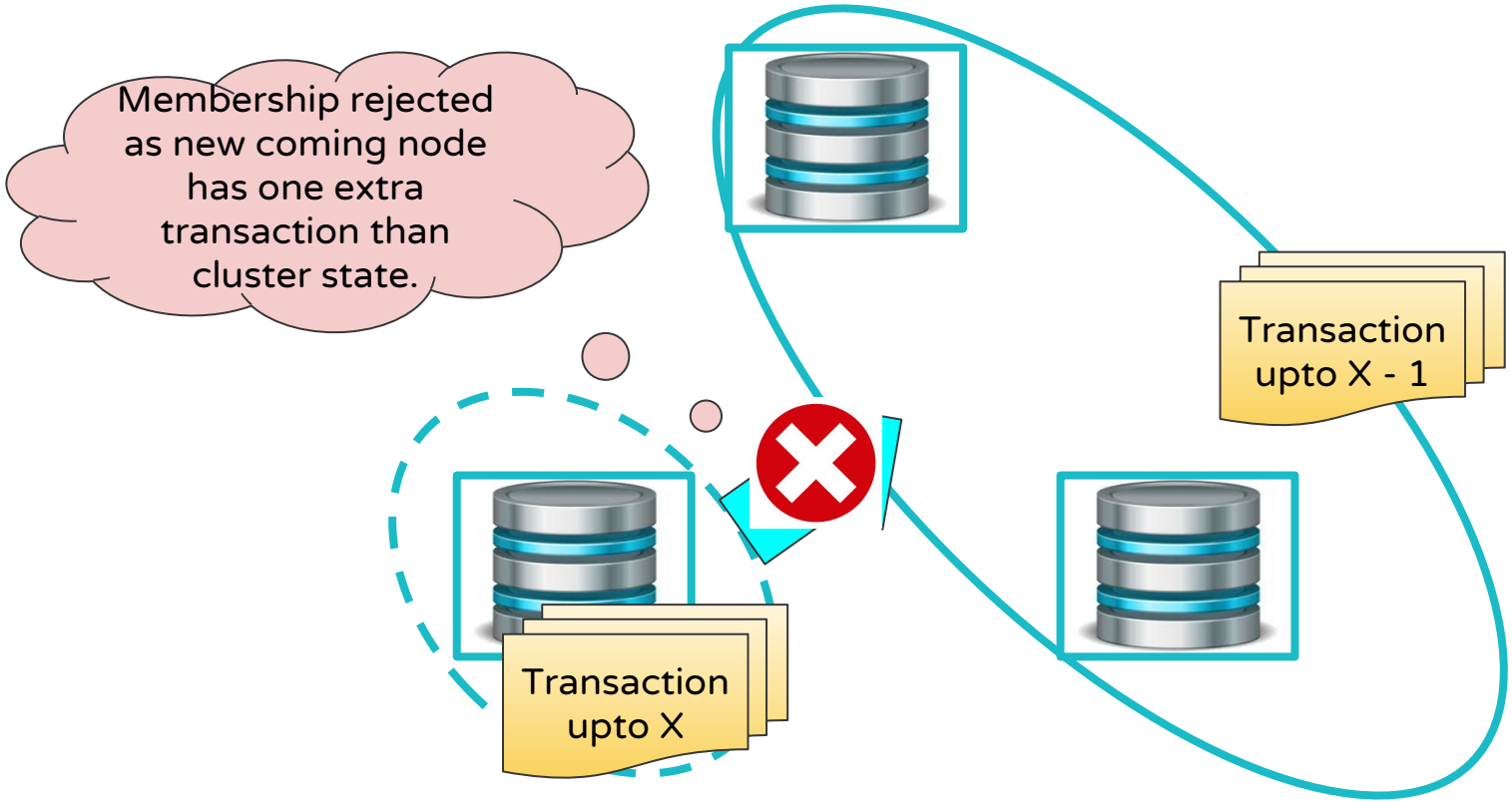
Scenario: Another aspect of data inconsistency



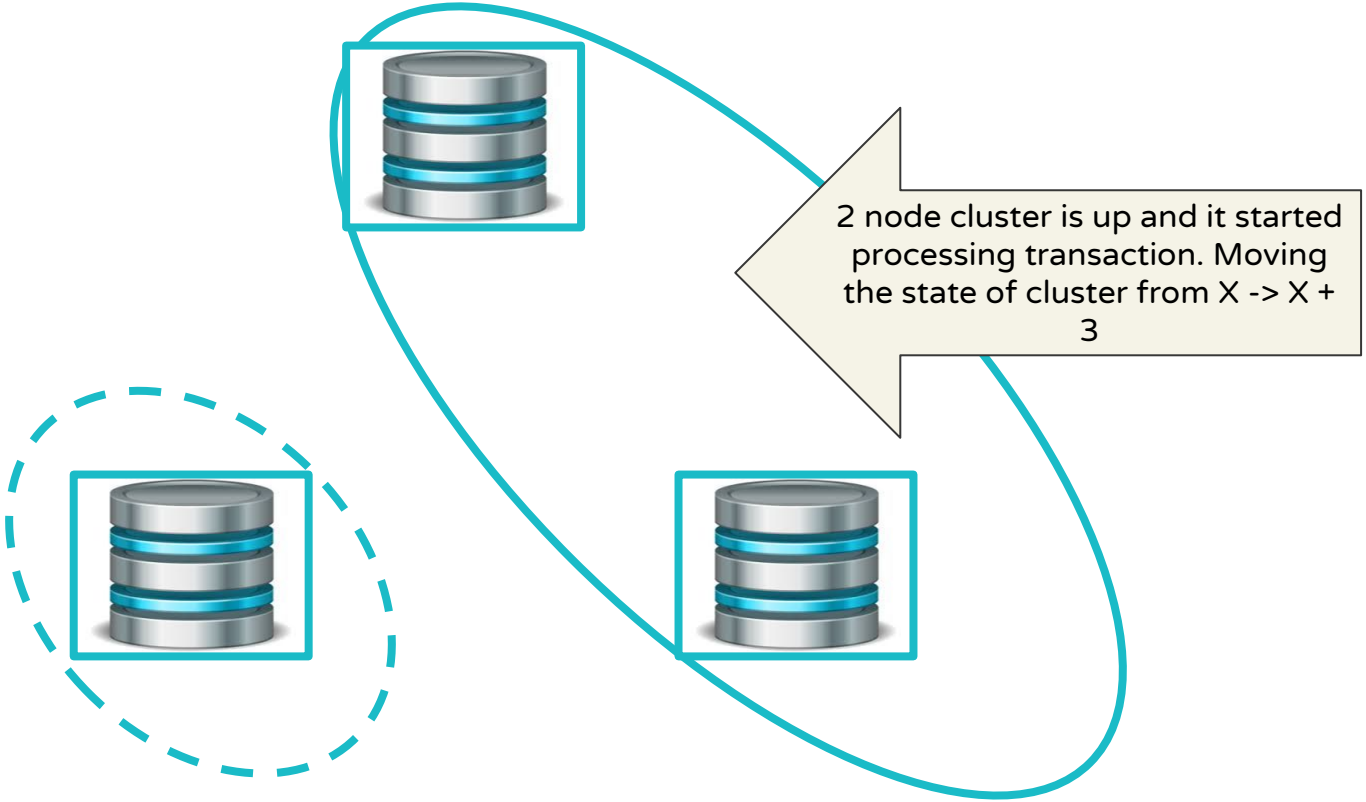
Scenario: Another aspect of data inconsistency



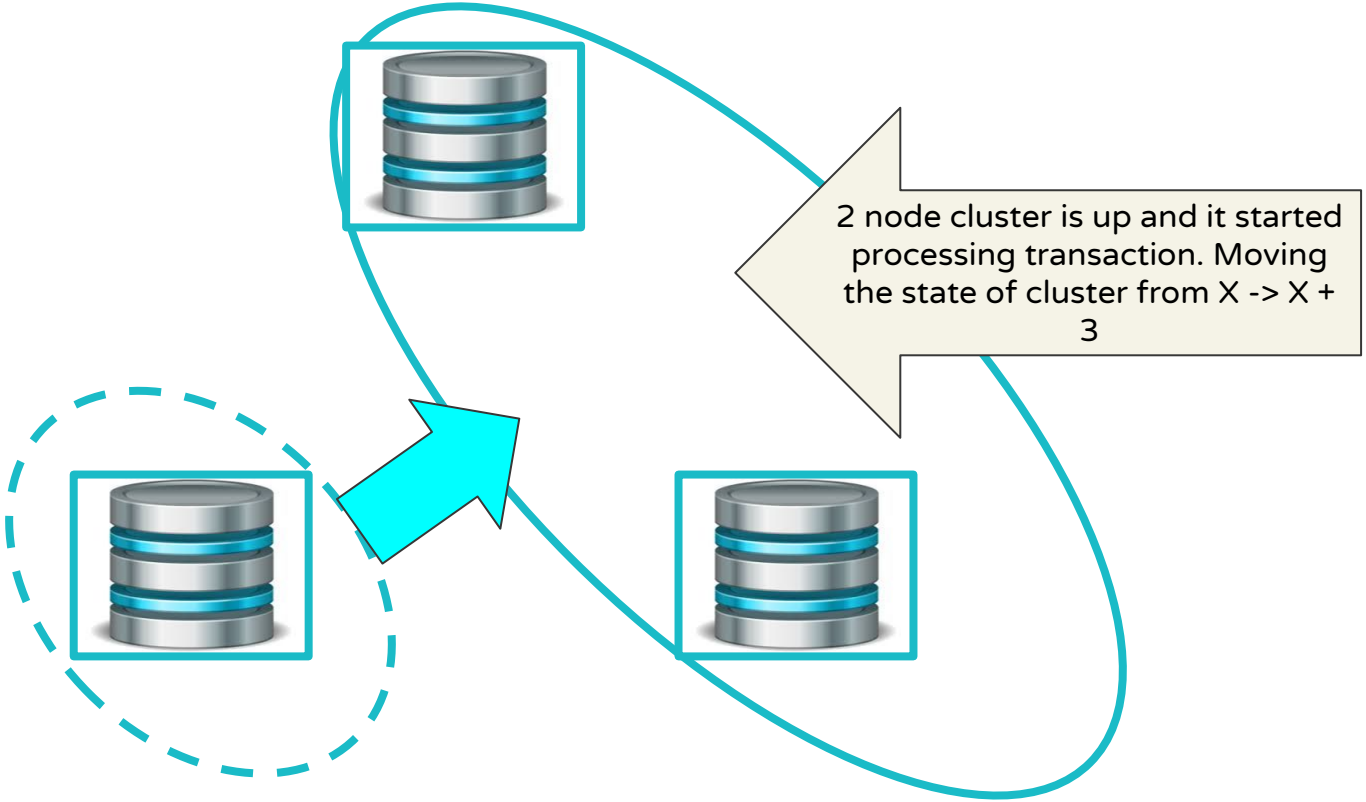
Scenario: Another aspect of data inconsistency



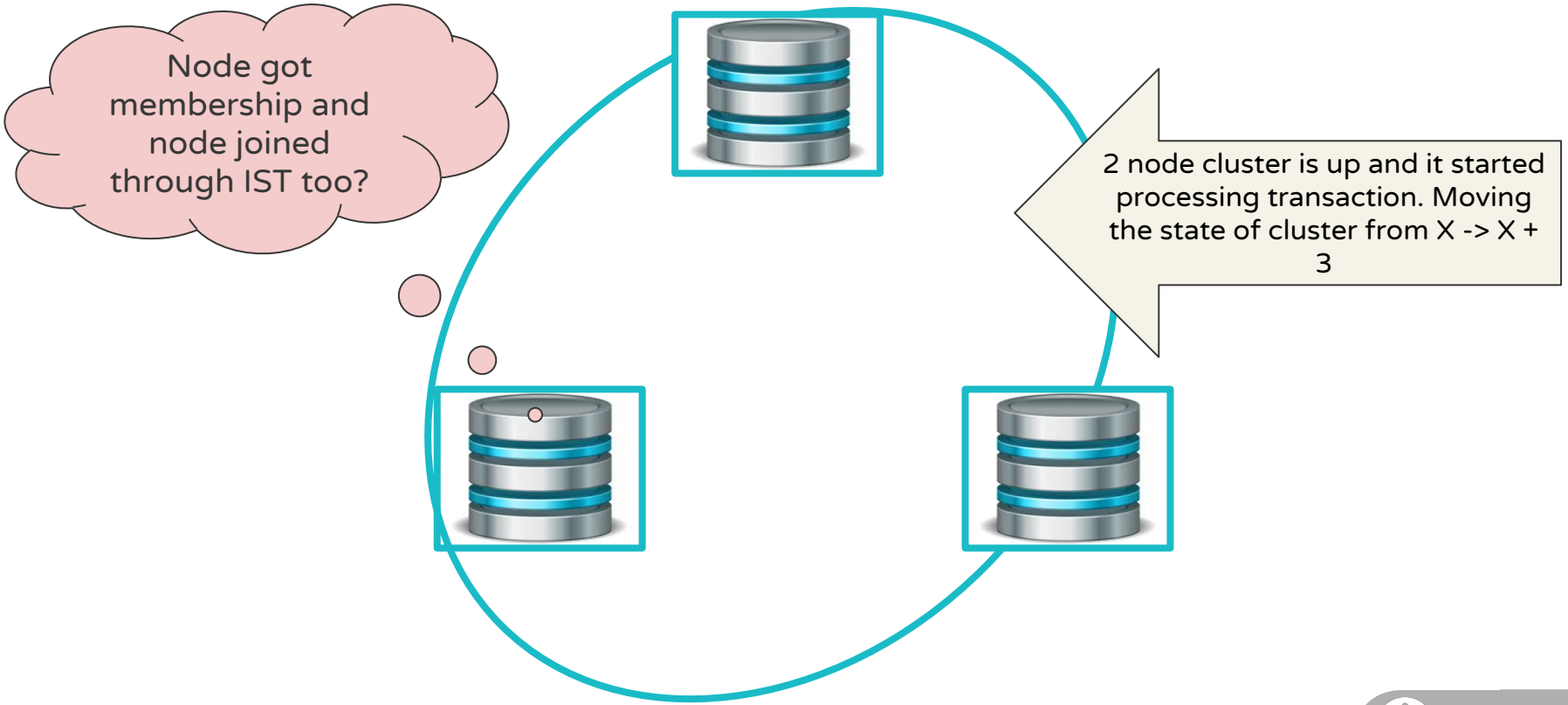
Scenario: Another aspect of data inconsistency



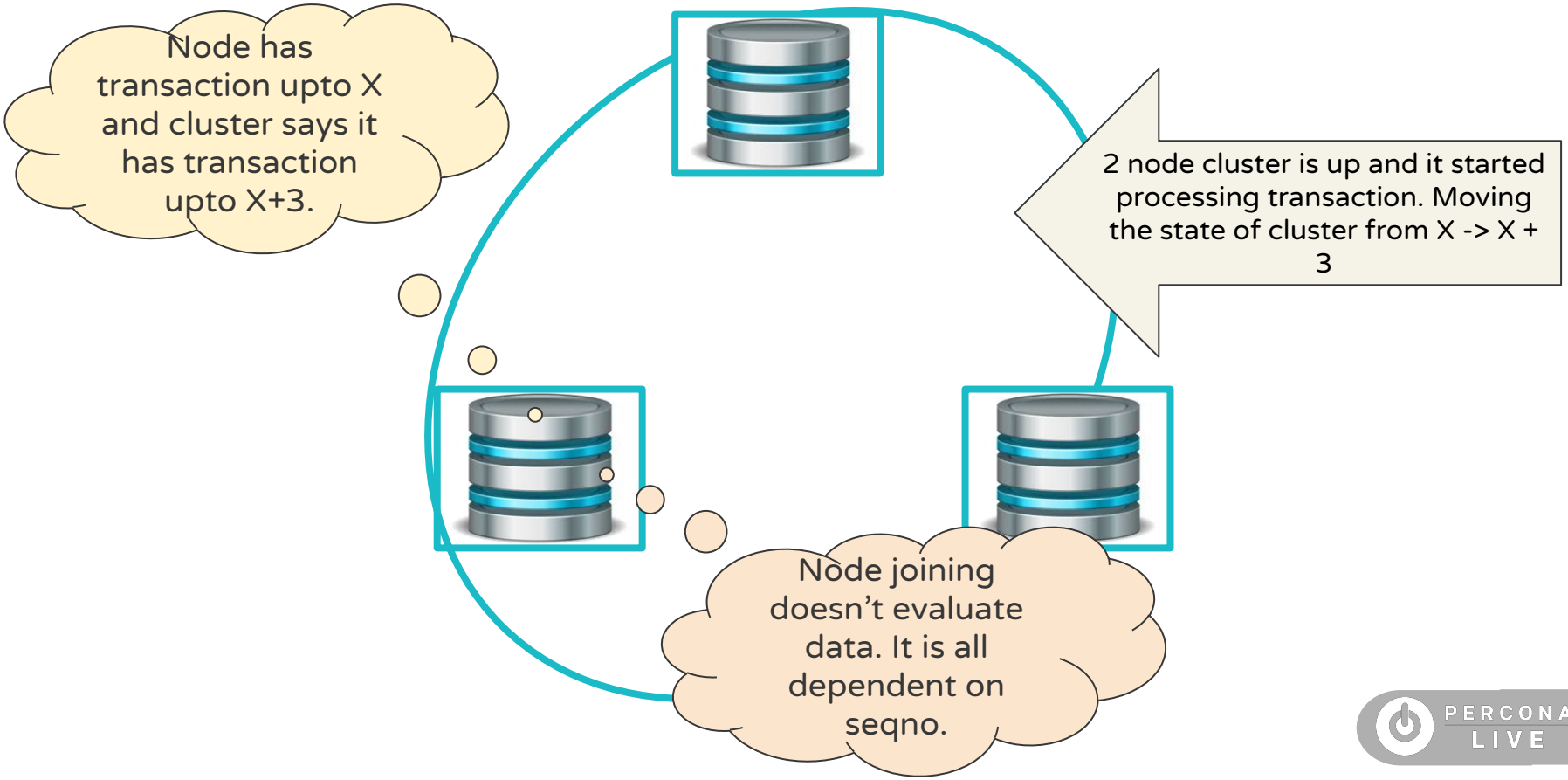
Scenario: Another aspect of data inconsistency



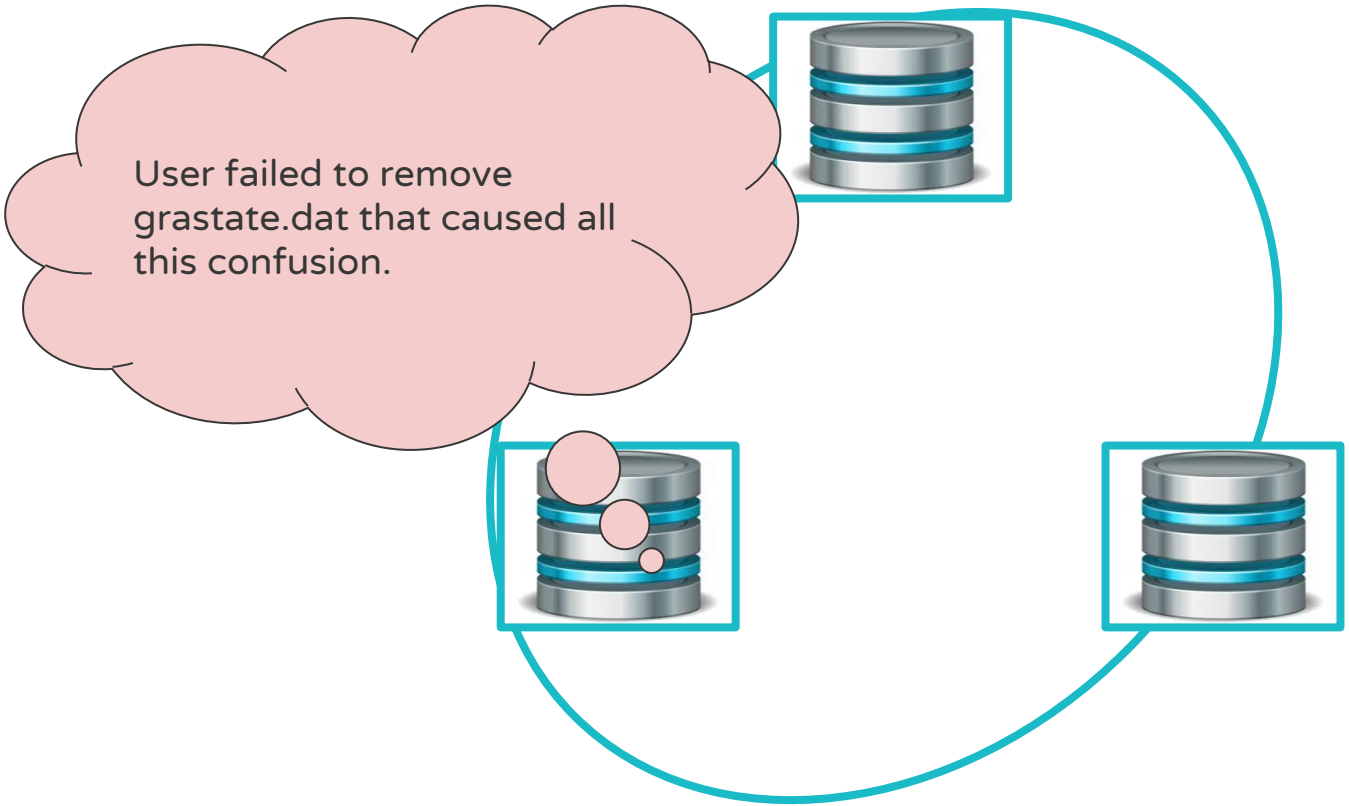
Scenario: Another aspect of data inconsistency



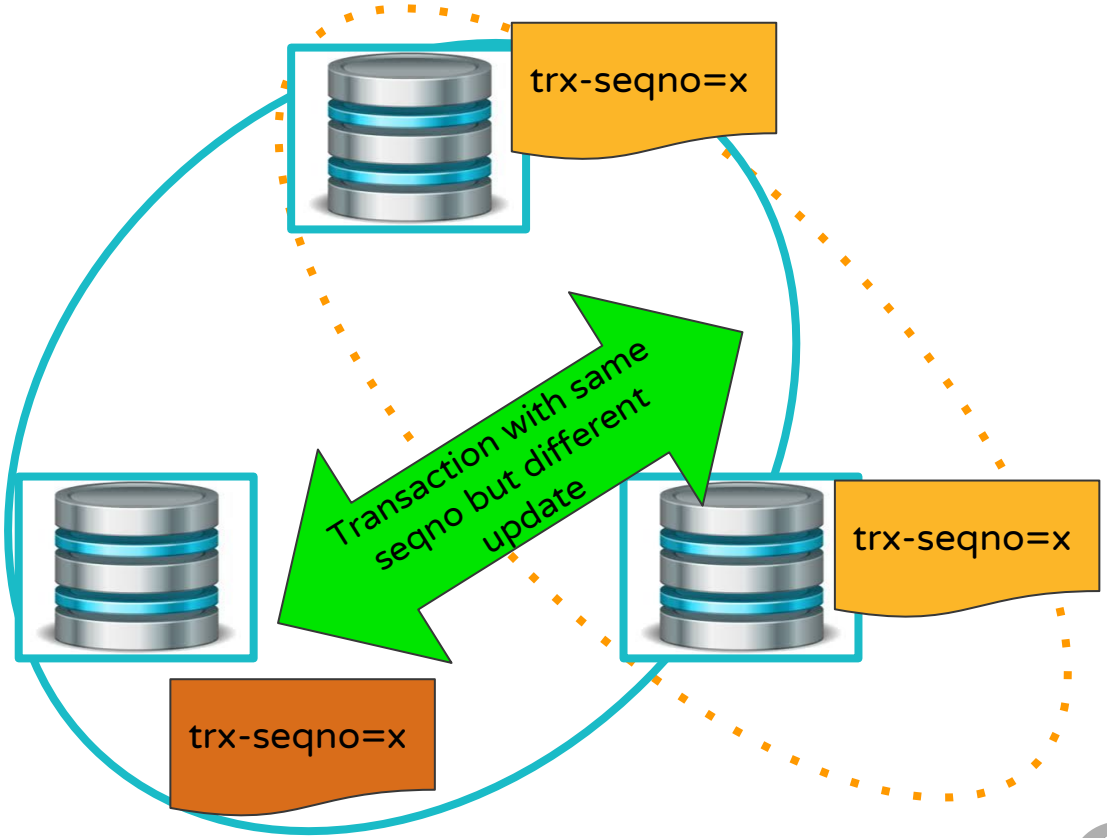
Scenario: Another aspect of data inconsistency



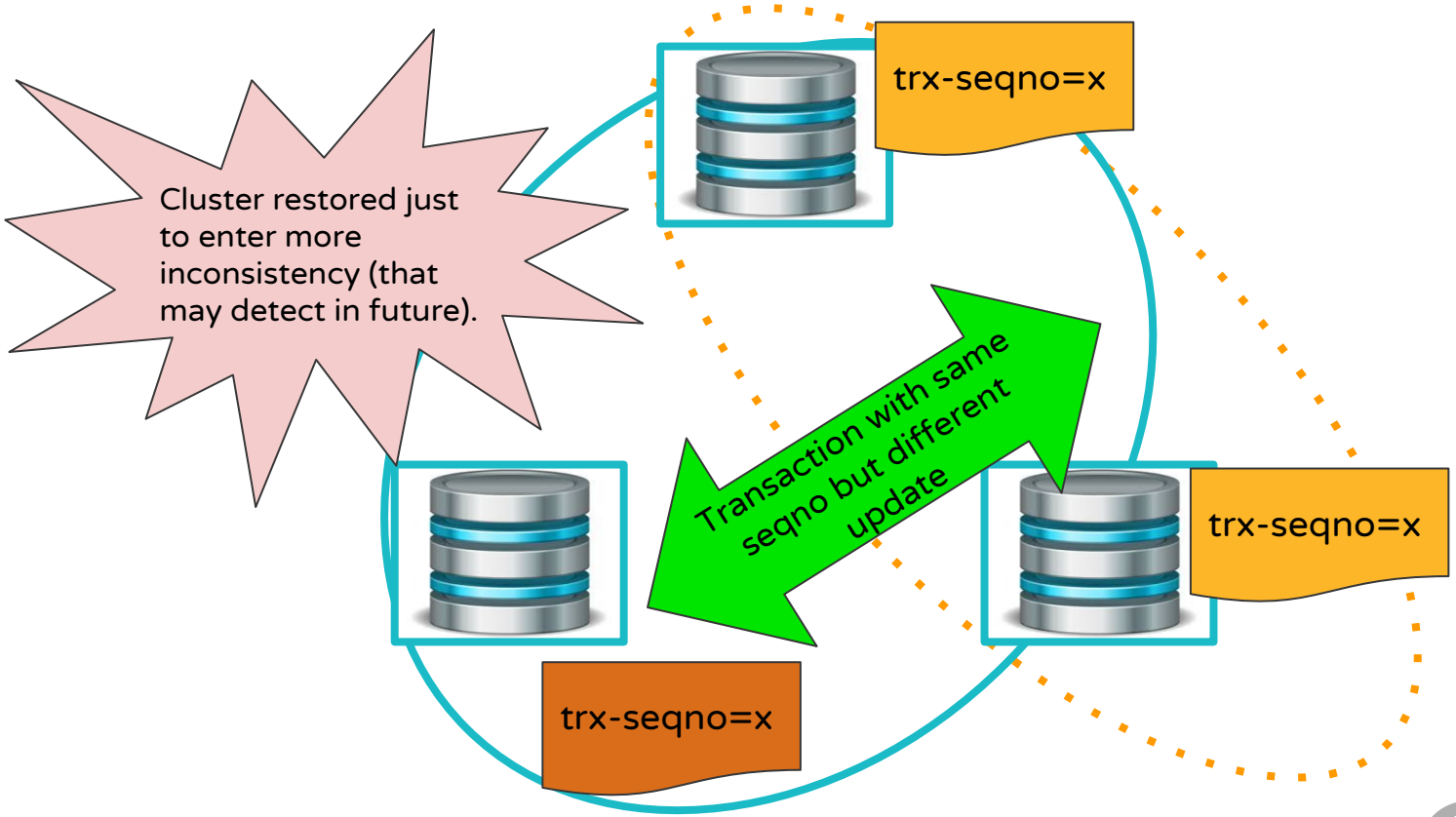
Scenario: Another aspect of data inconsistency



Scenario: Another aspect of data inconsistency



Scenario: Another aspect of data inconsistency



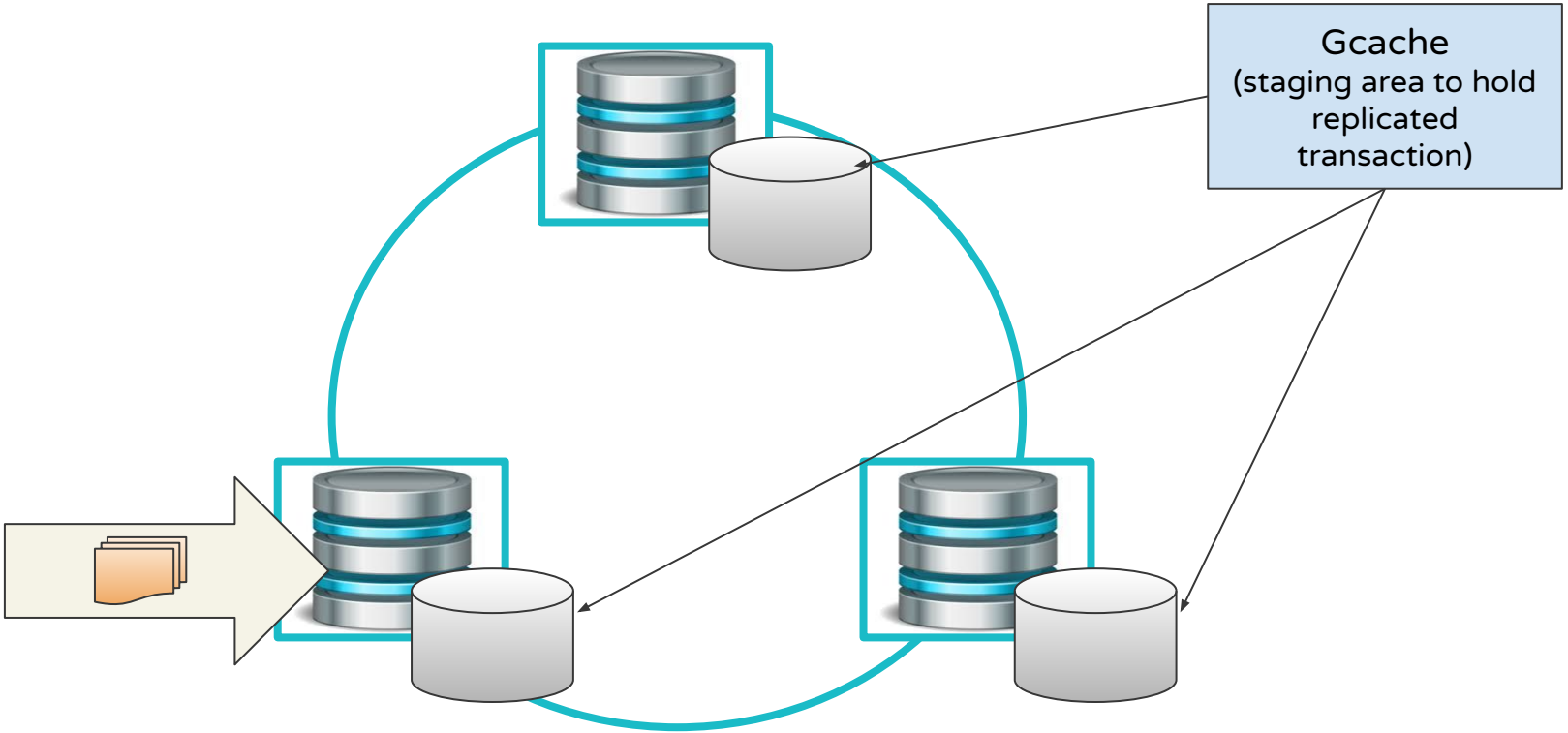
Scenario: Cluster doesn't come up on restart

Avoid running node local operation.

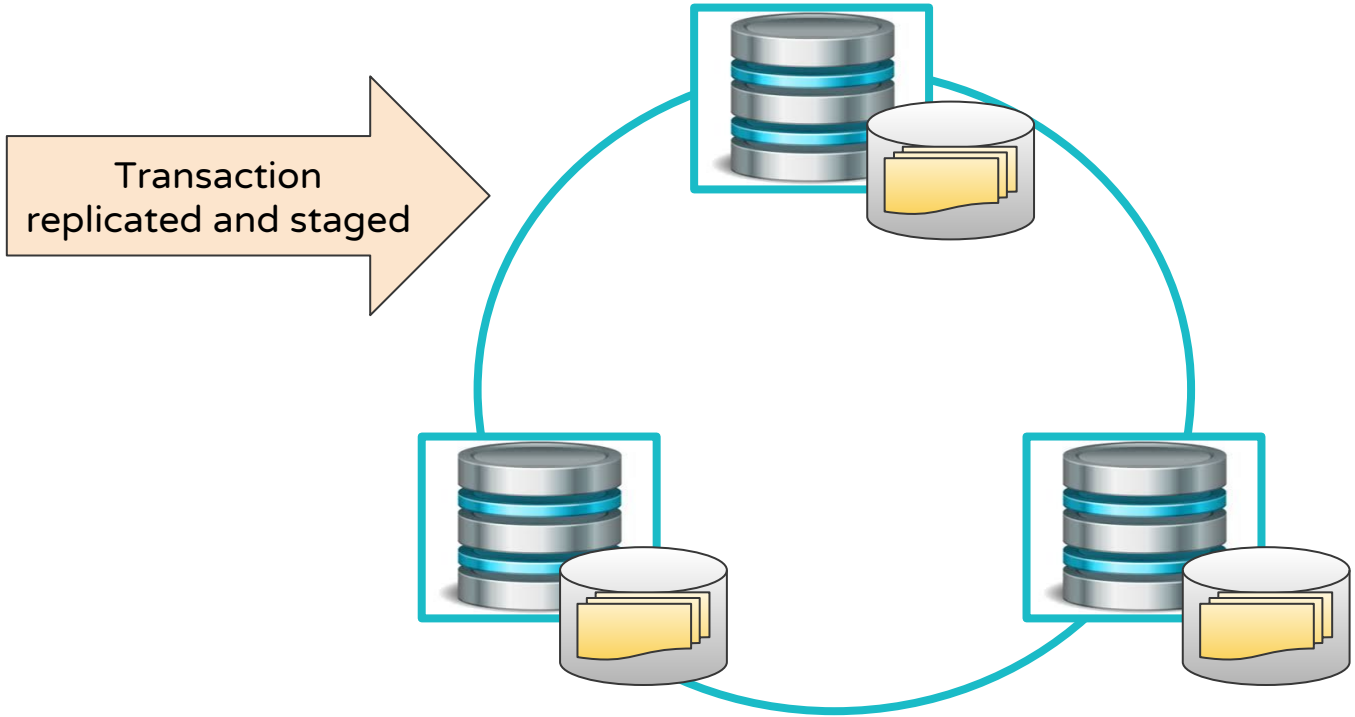
If cluster enter inconsistent state carefully follow the step-by-step guide to recover *(don't fear SST, it is for your good).*

Scenario: Delayed purging

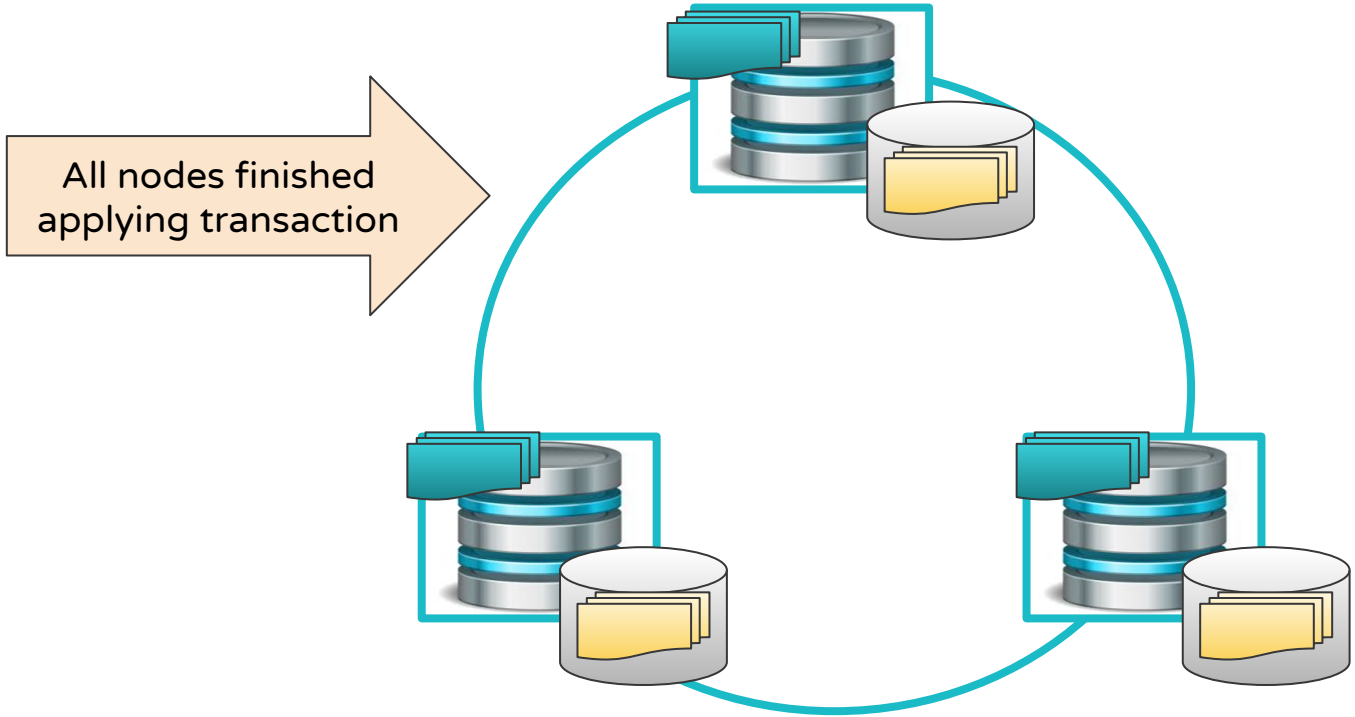
Scenario: Delayed purging



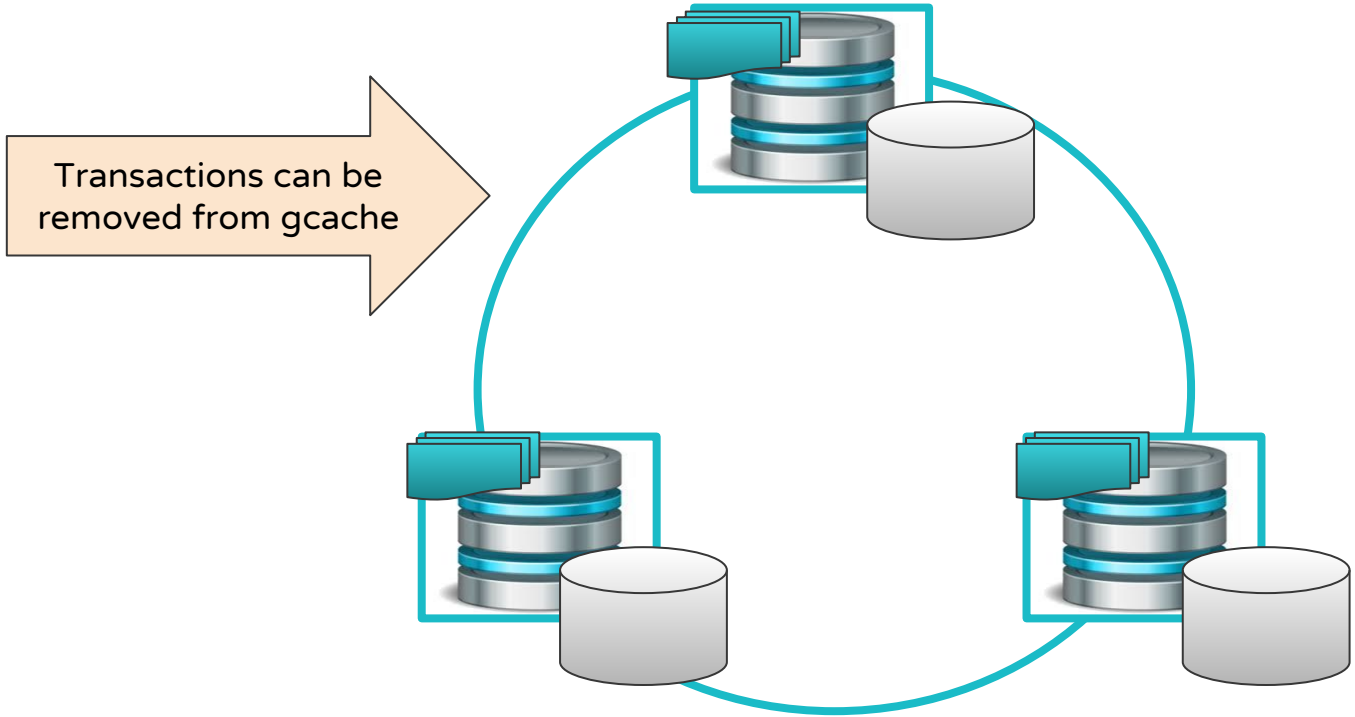
Scenario: Delayed purging



Scenario: Delayed purging



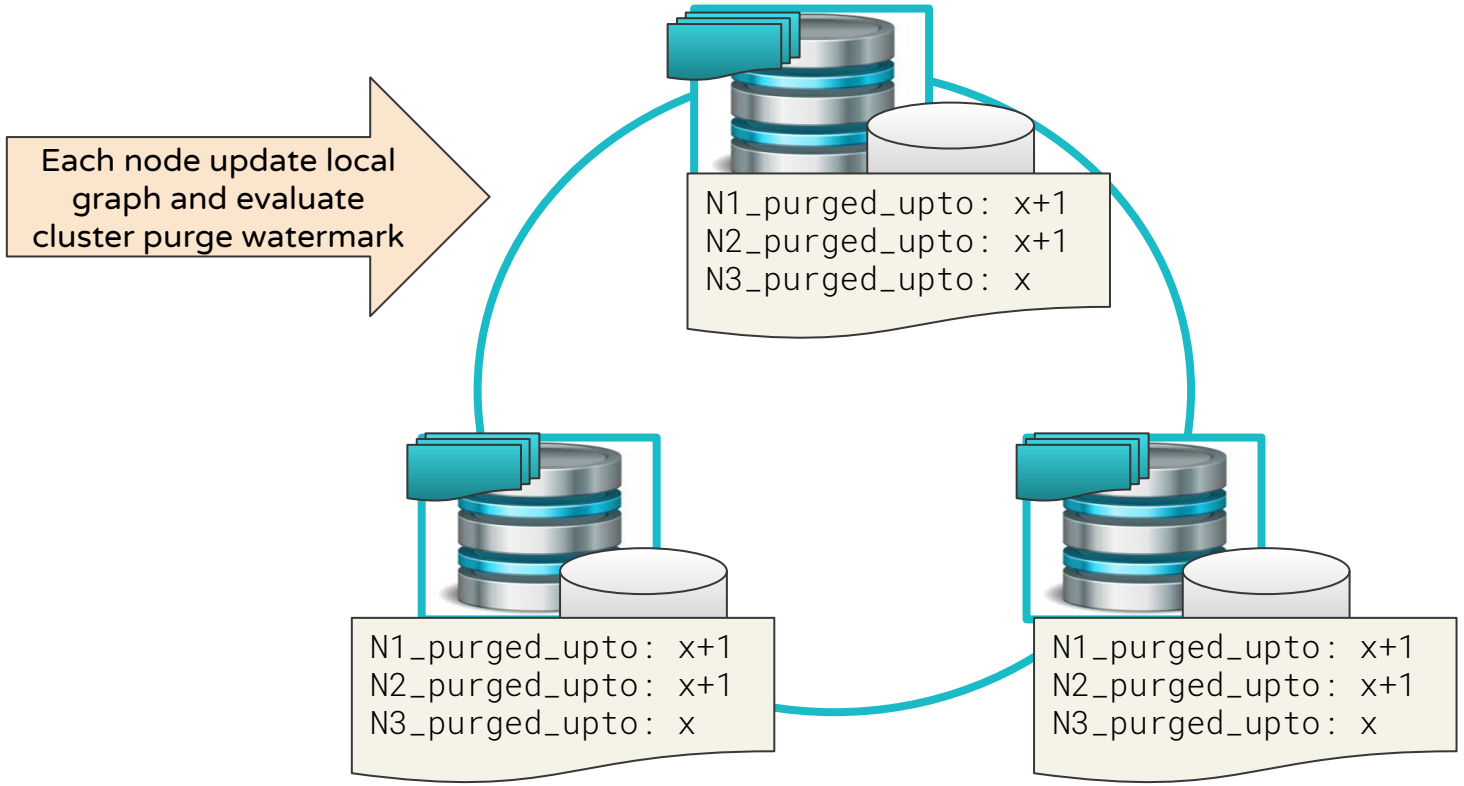
Scenario: Delayed purging



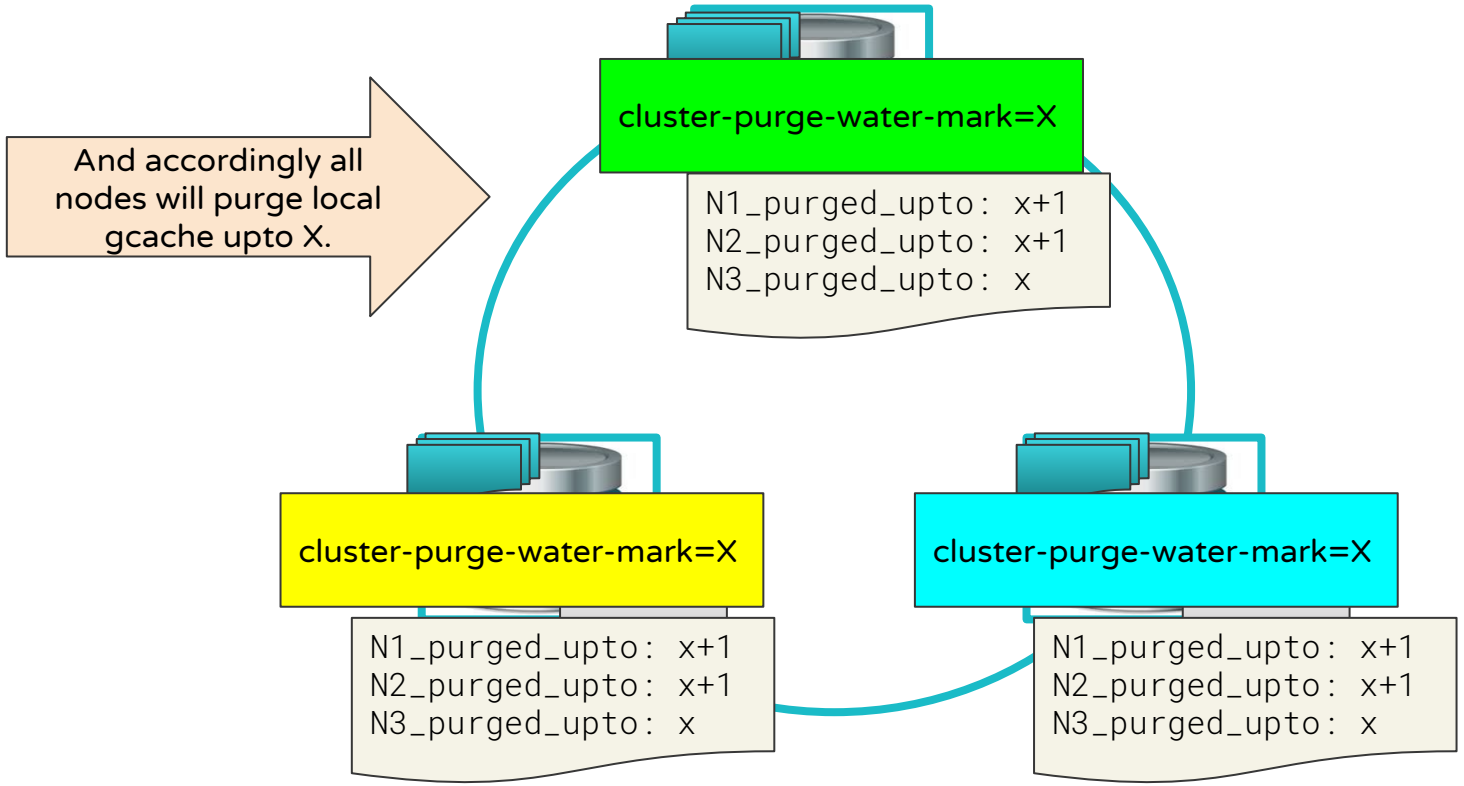
Scenario: Delayed purging

- Each node at configured interval notifies other nodes/cluster about its transaction committed status
- This configuration is controlled by 2 conditions:
  - `gcache.keep_page_size` and `gcache.keep_page_count`
  - `static limit on number of keys (1K), transactions (128), bytes (128M).`
- Accordingly each node evaluates the cluster level lowest water mark and initiate gcache purge.

Scenario: Delayed purging



Scenario: Delayed purging



Scenario: Delayed purging

```
2018-10-17T06:28:12.487964Z 0 [Note] WSREP: gcache: 1 page(s) deallocated...
2018-10-17T06:28:12.488070Z 0 [Note] WSREP: Deleted page /opt/projects/percona/merge/57-merge/installed/pxc-node/dn1/gcache.page.000034
2018-10-17T06:28:12.488212Z 0 [Note] WSREP: gcache: 1 page(s) deallocated...
2018-10-17T06:28:12.488370Z 0 [Note] WSREP: Deleted page /opt/projects/percona/merge/57-merge/installed/pxc-node/dn1/gcache.page.000035
2018-10-17T06:28:12.488460Z 0 [Note] WSREP: gcache: 1 page(s) deallocated...
2018-10-17T06:28:12.488600Z 0 [Note] WSREP: Deleted page /opt/projects/percona/merge/57-merge/installed/pxc-node/dn1/gcache.page.000036
2018-10-17T06:28:12.491245Z 0 [Note] WSREP: gcache: 1 page(s) deallocated...
2018-10-17T06:28:12.491405Z 0 [Note] WSREP: Deleted page /opt/projects/percona/merge/57-merge/installed/pxc-node/dn1/gcache.page.000037
2018-10-17T06:28:12.491486Z 0 [Note] WSREP: gcache: 1 page(s) deallocated...
2018-10-17T06:28:12.491632Z 0 [Note] WSREP: Deleted page /opt/projects/percona/merge/57-merge/installed/pxc-node/dn1/gcache.page.000038
2018-10-17T06:28:12.997919Z 0 [Note] WSREP: gcache: 1 page(s) deallocated...
2018-10-17T06:28:12.998078Z 0 [Note] WSREP: Deleted page /opt/projects/percona/merge/57-merge/installed/pxc-node/dn1/gcache.page.000039
```

gcache page created and purged.

Scenario: Delayed purging

```
2018-10-17T07:12:16.852599Z 0 [Note] [Debug] WSREP: gcs/src/gcs_group.cpp:gcs_group_handle_last_msg():650: New COMMIT CUT 2360 after 2360 from 1
2018-10-17T07:12:16.852705Z 4 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1033: purging index up to 2360
2018-10-17T07:12:16.854787Z 4 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1040: releasing seqno from gcache 2360
2018-10-17T07:12:16.854844Z 4 [Note] [Debug] WSREP: galera/src/replicator_smm.cpp:process_commit_cut():1464: Got commit cut from GCS: 2360
[kbauskar@kbxps installed] $
```

```
New COMMIT CUT 2360 after 2360 from 1
purging index up to 2360
releasing seqno from gcache 2360
Got commit cut from GCS: 2360
```

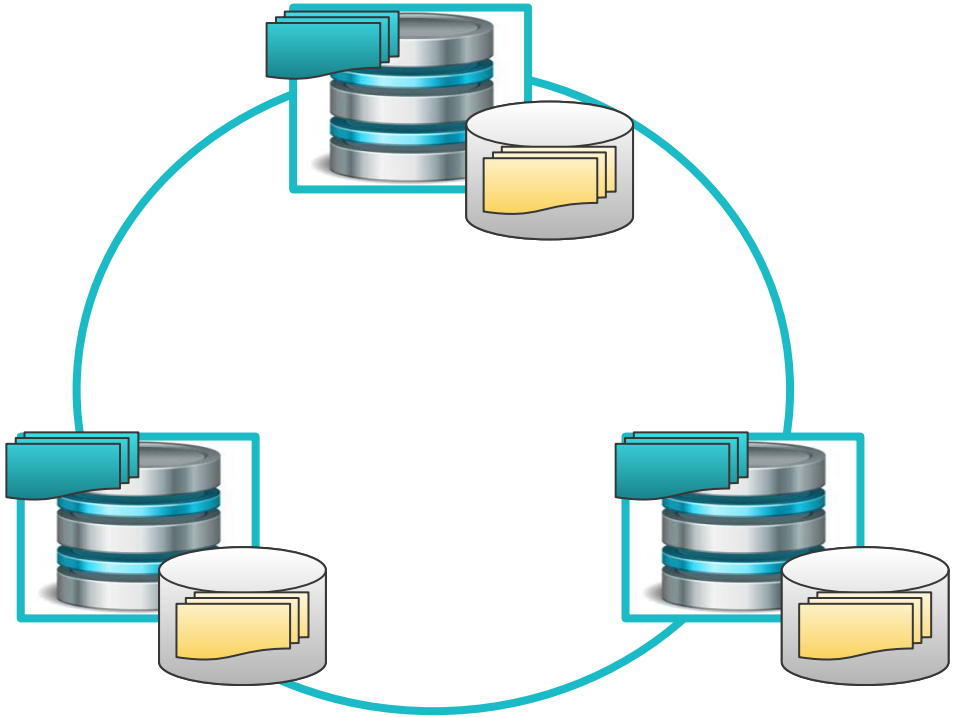
Scenario: Delayed purging

```
2018-10-17T07:12:16.7012072 18 [Note] [Debug] WSREP: galera/src/wsrep_provider.cpp:galera_post_rollback():472: trx 18446744073709551615 not found
2018-10-17T07:12:16.852599Z 0 [Note] [Debug] WSREP: gcs/src/gcs_group.cpp:gcs_group_handle_last_msg():650: New COMMIT CUT 2360 after 2360 from 1
2018-10-17T07:12:16.852705Z 4 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1033: purging index up to 2360
2018-10-17T07:12:16.854787Z 4 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1040: releasing seqno from gcache 2360
2018-10-17T07:12:16.854844Z 4 [Note] [Debug] WSREP: galera/src/replicator_smm.cpp:process_commit_cut():1464: Got commit cut from GCS: 2360
[kbauskar@kbxps installed] $
```

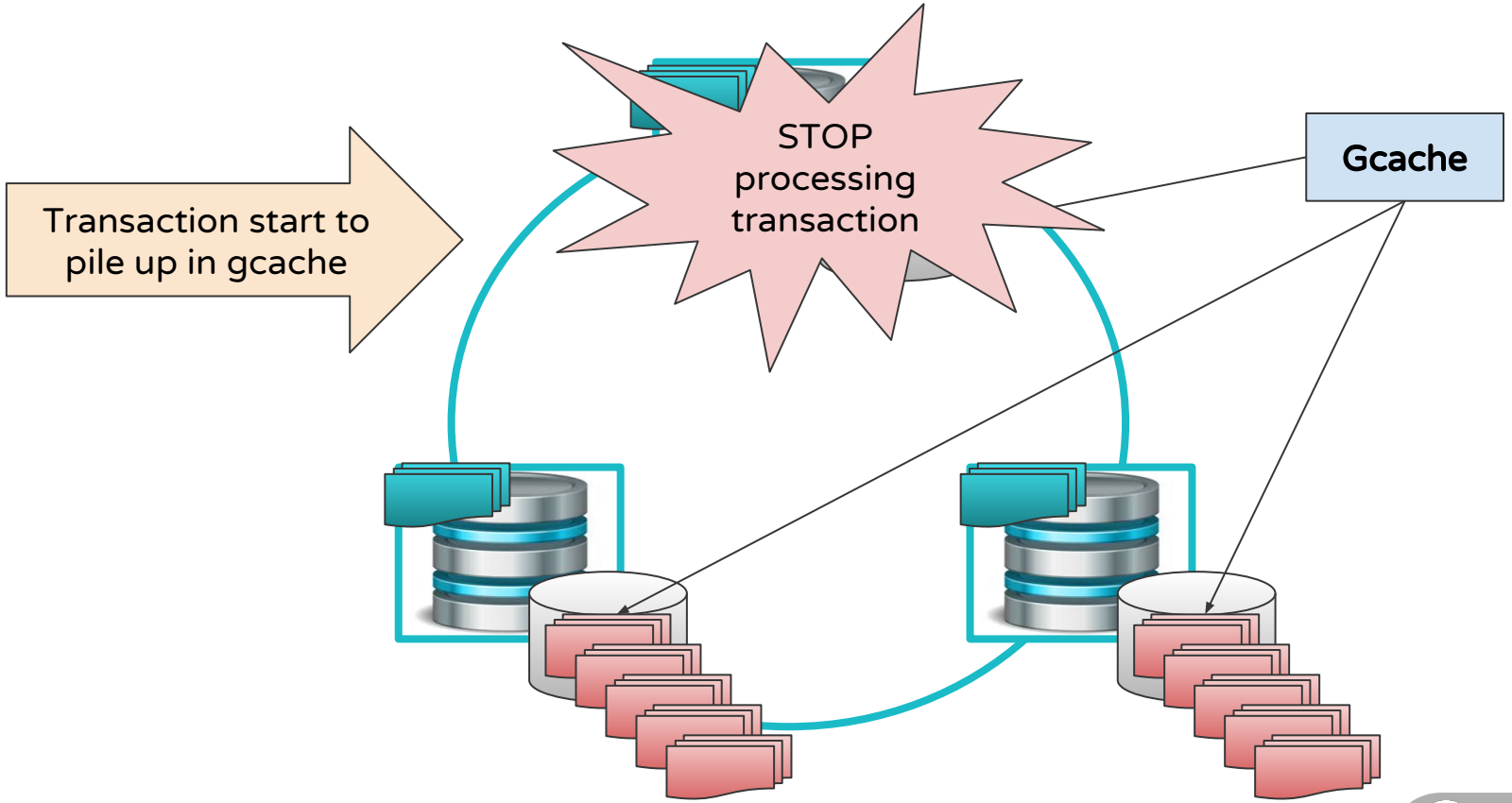
```
New COMMIT CUT 2360 after 2360 from 1
purging index up to 2360
releasing seqno from gcache 2360
Got commit cut from GCS: 2360
```

Regularly each node communicates, `committed upto` water mark and then as per protocol explained, purging initiates.

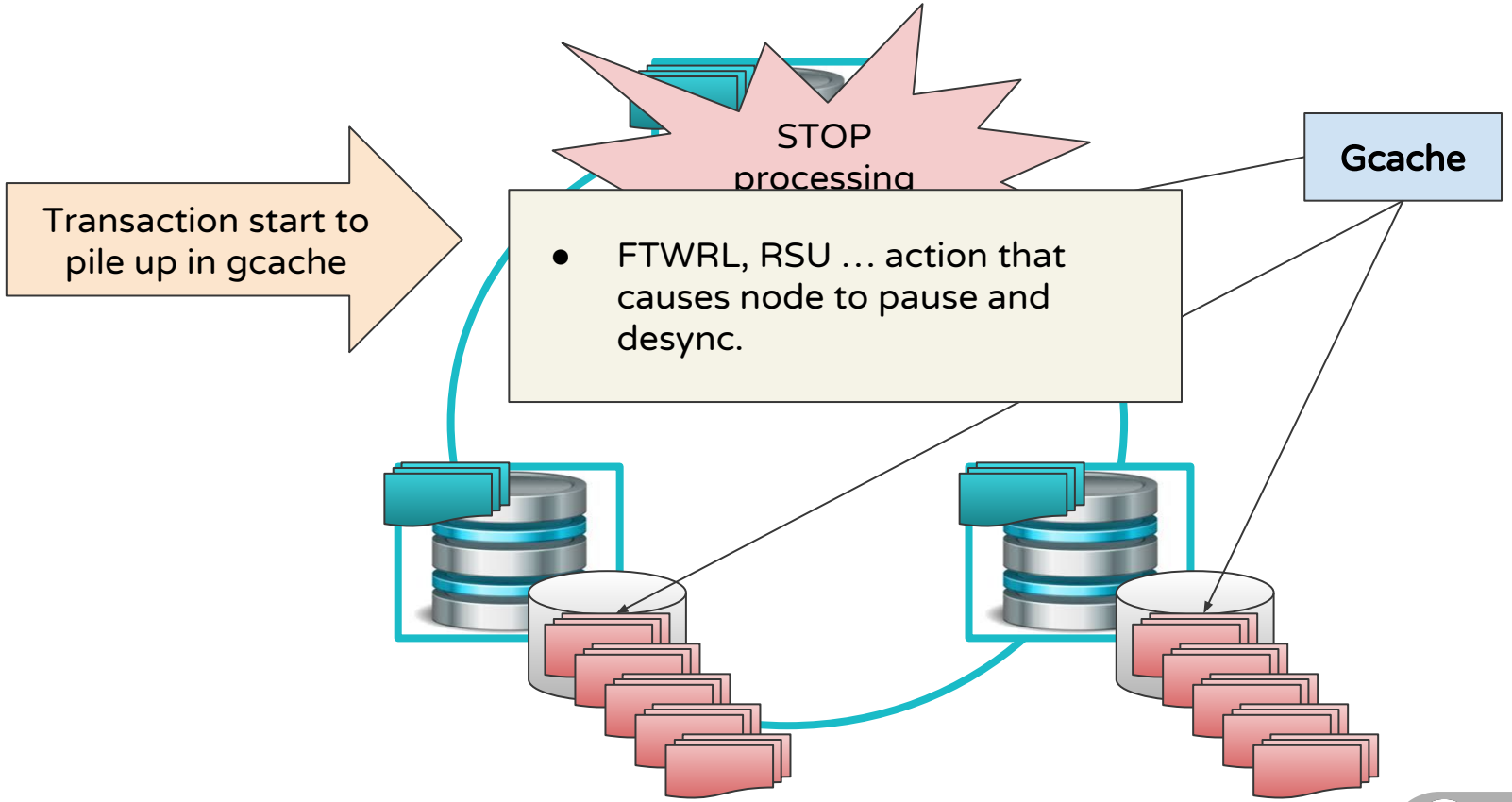
Scenario: Delayed purging



Scenario: Delayed purging



Scenario: Delayed purging



Scenario: Delayed purging

- Given that one of the node is not making progress it would not emit its transaction committed status.
- This would freeze the cluster-purge-water-mark as lowest transaction continue to lock-down.
- This means, though other nodes are making progress, they will continue to pile up galera cache.

Scenario: Delayed purging

- Given that one of the node is not making progress it would not emit its transaction committed status.
- This would freeze the cluster-purge-water-mark as lowest transaction continue to lock-down.
- This means, though other nodes are making progress, they will continue to pile up galera cache.

Galera has protection against it.
If number of transactions continue to grow beyond some hard limits it would force purge.

Scenario: Delayed purging

```
018-10-17T08:16:44.683080Z 22 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1040: releasing seqno from gcache 11008
018-10-17T08:16:44.943239Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:append_trx():1086: trx map size: 16511 - check if status.last_committed
is incrementing
018-10-17T08:16:44.943284Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1033: purging index up to 11136
018-10-17T08:16:44.944489Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1040: releasing seqno from gcache 11136
018-10-17T08:16:45.195568Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:append_trx():1086: trx map size: 16511 - check if status.last_committed
is incrementing
018-10-17T08:16:45.195591Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1033: purging index up to 11264
018-10-17T08:16:45.196374Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1040: releasing seqno from gcache 11264
```

```
trx map size: 16511 - check if status.last_committed is incrementing
purging index up to 11264
releasing seqno from gcache 11264
```

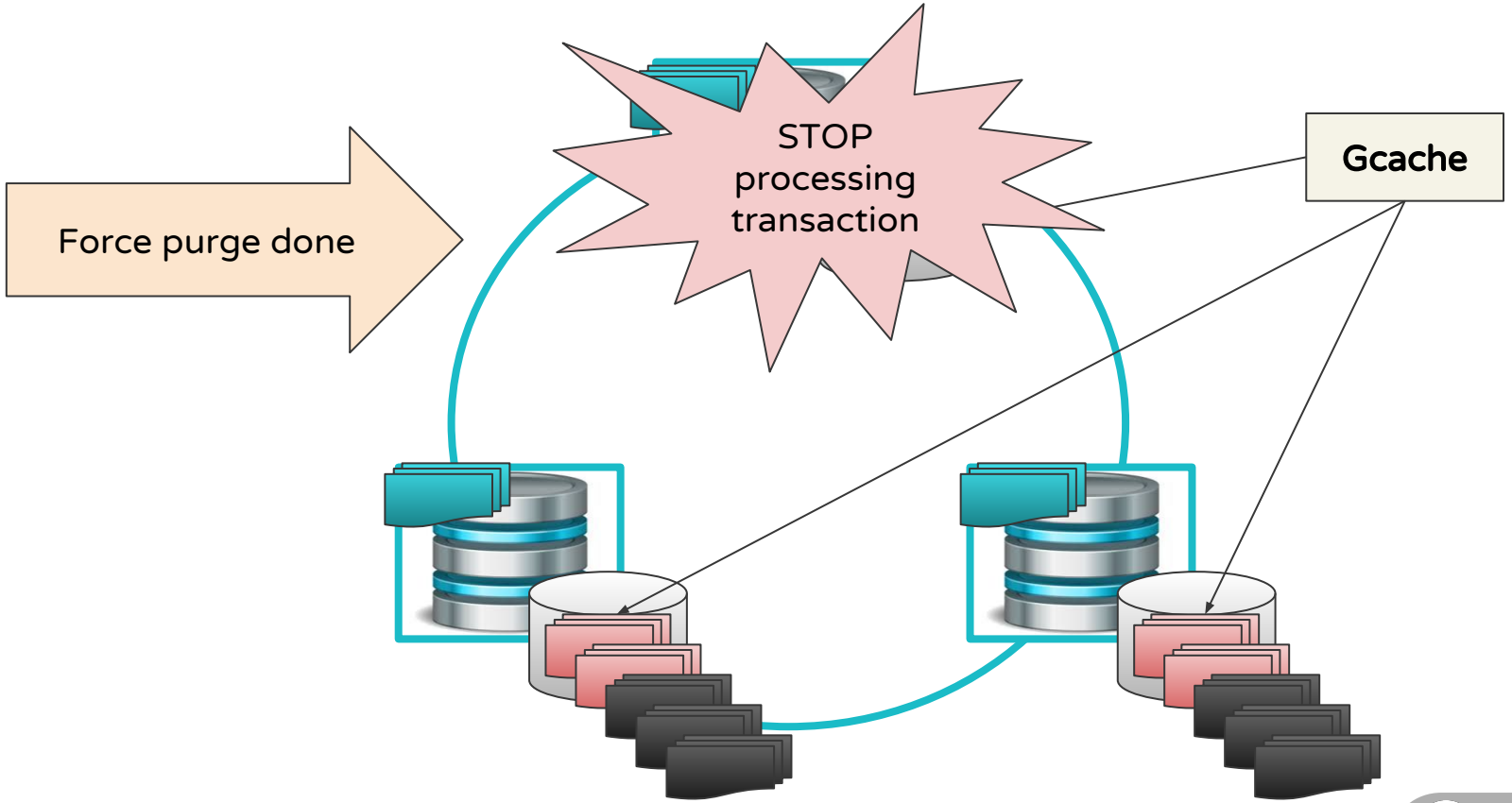
In-build mechanism to force purge.

Scenario: Delayed purging

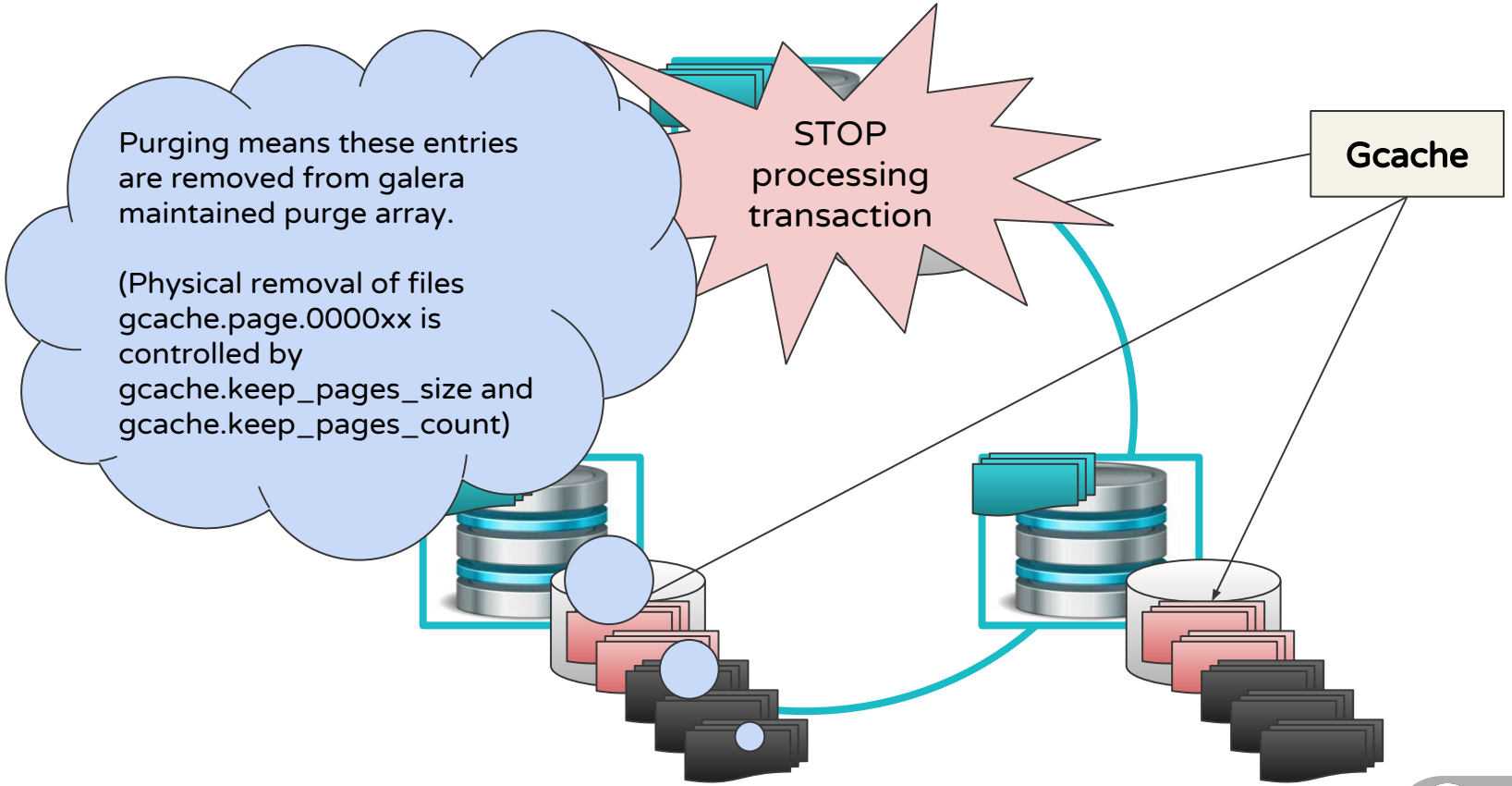
```
018-10-17T08:16:44.683080Z 22 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1040: releasing seqno from gcache 11008
018-10-17T08:16:44.943239Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:append_trx():1086: trx map size: 16511 - check if status.last_committed
is incrementing
018-10-17T08:16:44.943284Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1033: purging index up to 11136
018-10-17T08:16:44.944489Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1040: releasing seqno from gcache 11136
018-10-17T08:16:45.195568Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:append_trx():1086: trx map size: 16511 - check if status.last_committed
is incrementing
018-10-17T08:16:45.195591Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1033: purging index up to 11264
018-10-17T08:16:45.196374Z 23 [Note] [Debug] WSREP: galera/src/certification.cpp:purge_trxs_upto_():1040: releasing seqno from gcache 11264
```

```
trx map size: 16511 - check if status.last_committed is incrementing
purging index up to 11264
releasing seqno from gcache 11264
```

Scenario: Delayed purging



Scenario: Delayed purging



Scenario: Delayed purging

All nodes should have same configuration.

Keep a close watch if you plan to run a backup operation or other operation that can cause node to halt.

Monitor node is making progress by keeping watch on wsrep_last_applied/wsrep_last_committed.

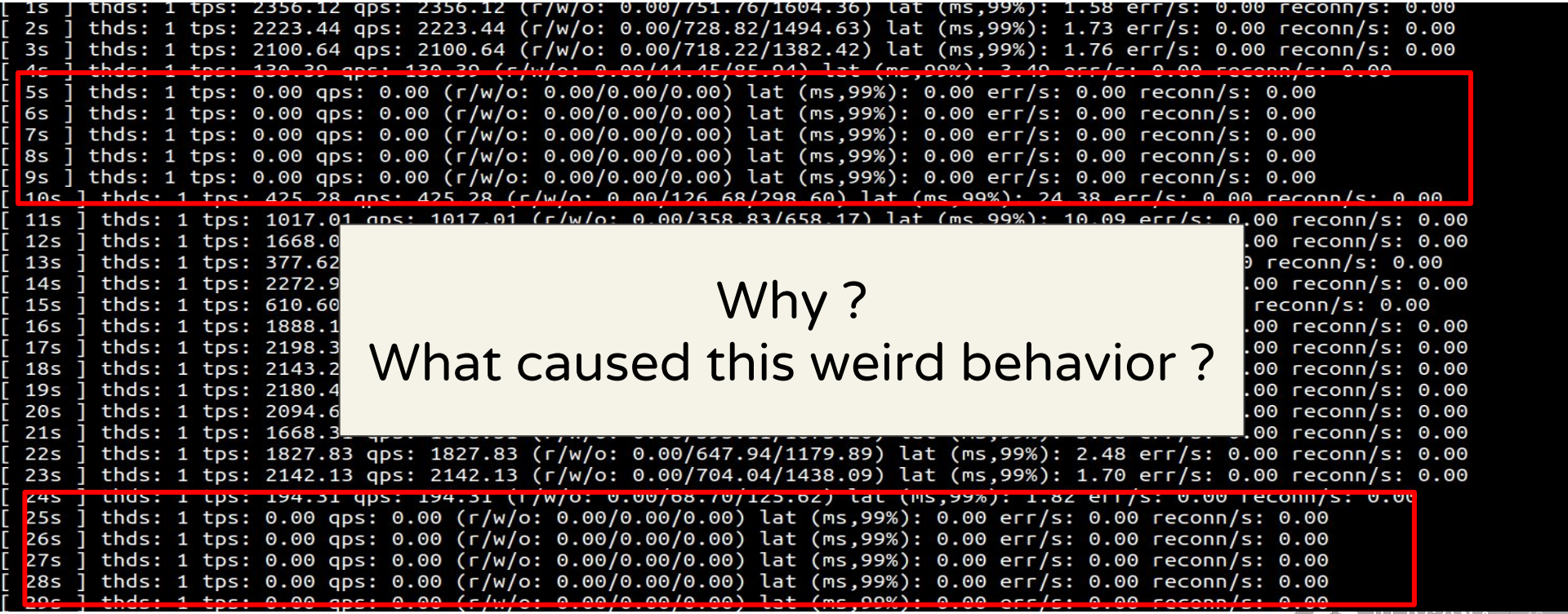
Scenario: Network latency and related failures

Scenario: Network latency and related failures

Scenario: Network latency and related failures

```
[ 1s ] thds: 1 tps: 2356.12 qps: 2356.12 (r/w/o: 0.00/751.76/1604.36) lat (ms,99%): 1.58 err/s: 0.00 reconn/s: 0.00
[ 2s ] thds: 1 tps: 2223.44 qps: 2223.44 (r/w/o: 0.00/728.82/1494.63) lat (ms,99%): 1.73 err/s: 0.00 reconn/s: 0.00
[ 3s ] thds: 1 tps: 2100.64 qps: 2100.64 (r/w/o: 0.00/718.22/1382.42) lat (ms,99%): 1.76 err/s: 0.00 reconn/s: 0.00
[ 4s ] thds: 1 tps: 130.39 qps: 130.39 (r/w/o: 0.00/44.45/85.94) lat (ms,99%): 3.49 err/s: 0.00 reconn/s: 0.00
[ 5s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 6s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 7s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 8s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 9s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 10s ] thds: 1 tps: 425.28 qps: 425.28 (r/w/o: 0.00/126.68/298.60) lat (ms,99%): 24.38 err/s: 0.00 reconn/s: 0.00
[ 11s ] thds: 1 tps: 1017.01 qps: 1017.01 (r/w/o: 0.00/358.83/658.17) lat (ms,99%): 10.09 err/s: 0.00 reconn/s: 0.00
[ 12s ] thds: 1 tps: 1668.05 qps: 1668.05 (r/w/o: 0.00/540.66/1127.39) lat (ms,99%): 4.03 err/s: 0.00 reconn/s: 0.00
[ 13s ] thds: 1 tps: 377.62 qps: 377.62 (r/w/o: 0.00/143.85/233.76) lat (ms,99%): 38.25 err/s: 0.00 reconn/s: 0.00
[ 14s ] thds: 1 tps: 2272.97 qps: 2272.97 (r/w/o: 0.00/752.97/1520.00) lat (ms,99%): 1.96 err/s: 0.00 reconn/s: 0.00
[ 15s ] thds: 1 tps: 610.60 qps: 610.60 (r/w/o: 0.00/213.51/397.09) lat (ms,99%): 1.82 err/s: 0.00 reconn/s: 0.00
[ 16s ] thds: 1 tps: 1888.19 qps: 1888.19 (r/w/o: 0.00/632.40/1255.79) lat (ms,99%): 2.14 err/s: 0.00 reconn/s: 0.00
[ 17s ] thds: 1 tps: 2198.33 qps: 2198.33 (r/w/o: 0.00/732.11/1466.22) lat (ms,99%): 1.52 err/s: 0.00 reconn/s: 0.00
[ 18s ] thds: 1 tps: 2143.21 qps: 2143.21 (r/w/o: 0.00/698.74/1444.47) lat (ms,99%): 1.76 err/s: 0.00 reconn/s: 0.00
[ 19s ] thds: 1 tps: 2180.45 qps: 2180.45 (r/w/o: 0.00/735.15/1445.30) lat (ms,99%): 1.55 err/s: 0.00 reconn/s: 0.00
[ 20s ] thds: 1 tps: 2094.64 qps: 2094.64 (r/w/o: 0.00/710.88/1383.76) lat (ms,99%): 1.76 err/s: 0.00 reconn/s: 0.00
[ 21s ] thds: 1 tps: 1668.31 qps: 1668.31 (r/w/o: 0.00/593.11/1075.20) lat (ms,99%): 3.68 err/s: 0.00 reconn/s: 0.00
[ 22s ] thds: 1 tps: 1827.83 qps: 1827.83 (r/w/o: 0.00/647.94/1179.89) lat (ms,99%): 2.48 err/s: 0.00 reconn/s: 0.00
[ 23s ] thds: 1 tps: 2142.13 qps: 2142.13 (r/w/o: 0.00/704.04/1438.09) lat (ms,99%): 1.70 err/s: 0.00 reconn/s: 0.00
[ 24s ] thds: 1 tps: 194.31 qps: 194.31 (r/w/o: 0.00/68.70/125.62) lat (ms,99%): 1.82 err/s: 0.00 reconn/s: 0.00
[ 25s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 26s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 27s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 28s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 29s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
```

Scenario: Network latency and related failures



Why ?

What caused this weird behavior ?

Scenario: Network latency and related failures

```
1s ] thds: 1 tps: 1004.40 qps: 1004.40 (r/w/o: 0.00/330.47/673.93) lat (ms,99%): 2.61 err/s: 0.00 reconn/s: 0.00
2s ] thds: 1 tps: 1798.95 qps: 1798.95 (r/w/o: 0.00/604.98/1193.96) lat (ms,99%): 2.22 err/s: 0.00 reconn/s: 0.00
3s ] thds: 1 tps: 1761.28 qps: 1761.28 (r/w/o: 0.00/546.09/1215.19) lat (ms,99%): 4.49 err/s: 0.00 reconn/s: 0.00
4s ] thds: 1 tps: 2205.29 qps: 2205.29 (r/w/o: 0.00/756.10/1449.19) lat (ms,99%): 1.58 err/s: 0.00 reconn/s: 0.00
5s ] thds: 1 tps: 2256.02 qps: 2256.02 (r/w/o: 0.00/764.01/1492.02) lat (ms,99%): 1.52 err/s: 0.00 reconn/s: 0.00
6s ] thds: 1 tps: 2204.72 qps: 2204.72 (r/w/o: 0.00/749.90/1454.81) lat (ms,99%): 1.64 err/s: 0.00 reconn/s: 0.00
7s ] thds: 1 tps: 2273.16 qps: 2273.16 (r/w/o: 0.00/716.05/1557.11) lat (ms,99%): 1.82 err/s: 0.00 reconn/s: 0.00
8s ] thds: 1 tps: 2232.03 qps: 2232.03 (r/w/o: 0.00/755.01/1477.02) lat (ms,99%): 1.58 err/s: 0.00 reconn/s: 0.00
9s ] thds: 1 tps: 2074.80 qps: 2074.80 (r/w/o: 0.00/701.93/1372.87) lat (ms,99%): 2.11 err/s: 0.00 reconn/s: 0.00
10s ] thds: 1 tps: 186.94 qps: 186.94 (r/w/o: 0.00/58.05/128.89) lat (ms,99%): 2.97 err/s: 0.00 reconn/s: 0.00
11s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
12s ] thds: 1 tps: 2.01 qps: 2.01 (r/w/o: 0.00/1.01/1.01) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
13s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
14s ] thds: 1 tps: 2.93 qps: 2.93 (r/w/o: 0.00/0.98/1.95) lat (ms,99%): 2320.55 err/s: 0.00 reconn/s: 0.00
15s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
16s ] thds: 1 tps: 1.00 qps: 1.00 (r/w/o: 0.00/1.00/0.00) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
17s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
18s ] thds: 1 tps: 1.00 qps: 1.00 (r/w/o: 0.00/1.00/0.00) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
19s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
20s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
21s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
22s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
23s ] thds: 1 tps: 1384.62 qps: 1384.62 (r/w/o: 0.00/472.99/911.62) lat (ms,99%): 2.26 err/s: 0.00 reconn/s: 0.00
24s ] thds: 1 tps: 2224.98 qps: 2224.98 (r/w/o: 0.00/719.29/1505.69) lat (ms,99%): 2.07 err/s: 0.00 reconn/s: 0.00
25s ] thds: 1 tps: 2061.52 qps: 2061.52 (r/w/o: 0.00/690.84/1370.68) lat (ms,99%): 1.86 err/s: 0.00 reconn/s: 0.00
26s ] thds: 1 tps: 1814.65 qps: 1814.65 (r/w/o: 0.00/590.21/1224.44) lat (ms,99%): 2.22 err/s: 0.00 reconn/s: 0.00
27s ] thds: 1 tps: 1794.67 qps: 1794.67 (r/w/o: 0.00/587.89/1206.78) lat (ms,99%): 2.18 err/s: 0.00 reconn/s: 0.00
28s ] thds: 1 tps: 2086.06 qps: 2086.06 (r/w/o: 0.00/669.02/1417.04) lat (ms,99%): 1.67 err/s: 0.00 reconn/s: 0.00
```

Scenario: Network latency and related failures

```
1s ] thds: 1 tps: 1004.40 qps: 1004.40 (r/w/o: 0.00/330
2s ] thds: 1 tps: 1798.95 qps: 1798.95 (r/w/o: 0.00/604
3s ] thds: 1 tps: 1761.28 qps: 1761.28 (r/w/o: 0.00/546
4s ] thds: 1 tps: 2205.29 qps: 2205.29 (r/w/o: 0.00/756
5s ] thds: 1 tps: 2256.02 qps: 2256.02 (r/w/o: 0.00/764
6s ] thds: 1 tps: 2204.72 qps: 2204.72 (r/w/o: 0.00/749
7s ] thds: 1 tps: 2273.16 qps: 2273.16 (r/w/o: 0.00/716
8s ] thds: 1 tps: 2232.03 qps: 2232.03 (r/w/o: 0.00/755
9s ] thds: 1 tps: 2074.80 qps: 2074.80 (r/w/o: 0.00/701
10s ] thds: 1 tps: 186.94 qps: 186.94 (r/w/o: 0.00/58.65/128.29) lat (ms,99%): 2.97 err/s: 0.00 reconn/s: 0.00
11s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
12s ] thds: 1 tps: 2.01 qps: 2.01 (r/w/o: 0.00/1.01/1.01) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
13s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
14s ] thds: 1 tps: 2.93 qps: 2.93 (r/w/o: 0.00/0.98/1.95) lat (ms,99%): 2320.55 err/s: 0.00 reconn/s: 0.00
15s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
16s ] thds: 1 tps: 1.00 qps: 1.00 (r/w/o: 0.00/1.00/0.00) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
17s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
18s ] thds: 1 tps: 1.00 qps: 1.00 (r/w/o: 0.00/1.00/0.00) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
19s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
20s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
21s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
22s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
23s ] thds: 1 tps: 1384.62 qps: 1384.62 (r/w/o: 0.00/472.99/911.62) lat (ms,99%): 2.26 err/s: 0.00 reconn/s: 0.00
24s ] thds: 1 tps: 2224.98 qps: 2224.98 (r/w/o: 0.00/719.29/1505.69) lat (ms,99%): 2.07 err/s: 0.00 reconn/s: 0.00
25s ] thds: 1 tps: 2061.52 qps: 2061.52 (r/w/o: 0.00/690.84/1370.68) lat (ms,99%): 1.86 err/s: 0.00 reconn/s: 0.00
26s ] thds: 1 tps: 1814.65 qps: 1814.65 (r/w/o: 0.00/590.21/1224.44) lat (ms,99%): 2.22 err/s: 0.00 reconn/s: 0.00
27s ] thds: 1 tps: 1794.67 qps: 1794.67 (r/w/o: 0.00/587.89/1206.78) lat (ms,99%): 2.18 err/s: 0.00 reconn/s: 0.00
28s ] thds: 1 tps: 2086.06 qps: 2086.06 (r/w/o: 0.00/669.02/1417.04) lat (ms,99%): 1.67 err/s: 0.00 reconn/s: 0.00
```

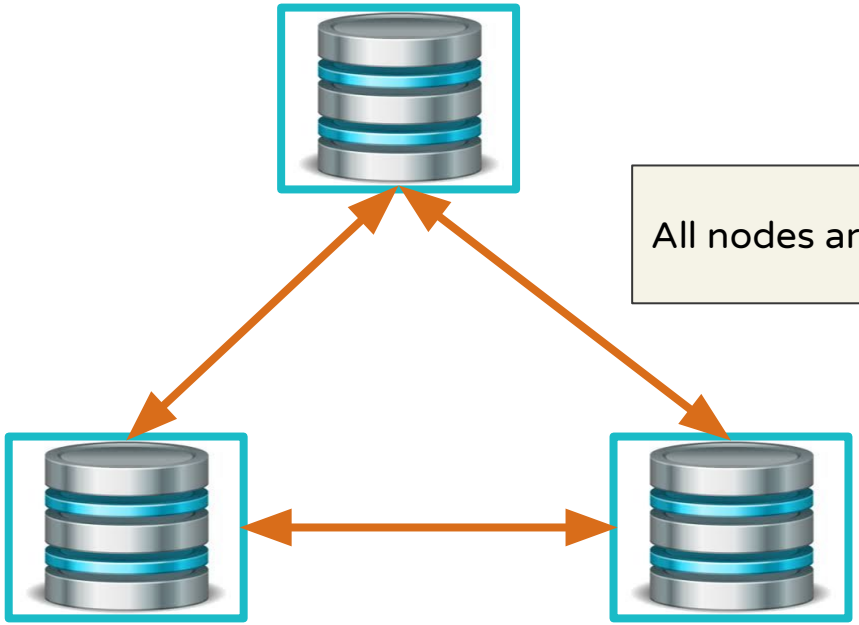
Cluster is neither complete down nor complete up. What's going on ? What is causing this weird behavior ?

Scenario: Network latency and related failures

```
2018-10-22T05:28:49.380284Z 7 [Note] WSREP: New cluster view: global state: c0ef3ed5-d1c2-11e8-b7b3-77974da4d118:1091590, view# 14: Primary, number of no
des: 3, my index: 0, protocol version 3
2018-10-22T05:28:49.380287Z 7 [Note] WSREP: Setting wsrep_ready to true
2018-10-22T05:28:49.380289Z 7 [Note] WSREP: Auto Increment Offset/Increment re-align with cluster membership change (Offset: 1 -> 1) (Increment: 3 -> 3)
2018-10-22T05:28:49.380292Z 7 [Note] WSREP: wsrep_notify_cmd is not defined, skipping notification.
2018-10-22T05:28:49.380400Z 7 [Note] WSREP: Assign initial position for certification: 1091590, protocol version: 4
2018-10-22T05:28:49.380553Z 18 [Note] WSREP: --------- CONFLICT DETECTED --------
2018-10-22T05:28:49.380563Z 18 [Note] WSREP: cluster conflict due to certification failure for threads:

2018-10-22T05:28:49.380566Z 18 [Note] WSREP: Victim thread:
   THD: 18, mode: local, state: executing, conflict: cert failure, seqno: -1
   SQL: UPDATE sbtest4 SET c=? WHERE id=?

2018-10-22T05:28:49.380864Z 18 [Note] WSREP: --------- CONFLICT DETECTED --------
2018-10-22T05:28:49.380930Z 18 [Note] WSREP: cluster conflict due to certification failure for threads:

2018-10-22T05:28:49.380934Z 18 [Note] WSREP: Victim thread:
   THD: 18, mode: local, state: executing, conflict: cert failure, seqno: -1
   SQL: UPDATE sbtest2 SET c=? WHERE id=?

2018-10-22T05:28:49.381183Z 0 [Note] WSREP: Service thread queue flushed.
```

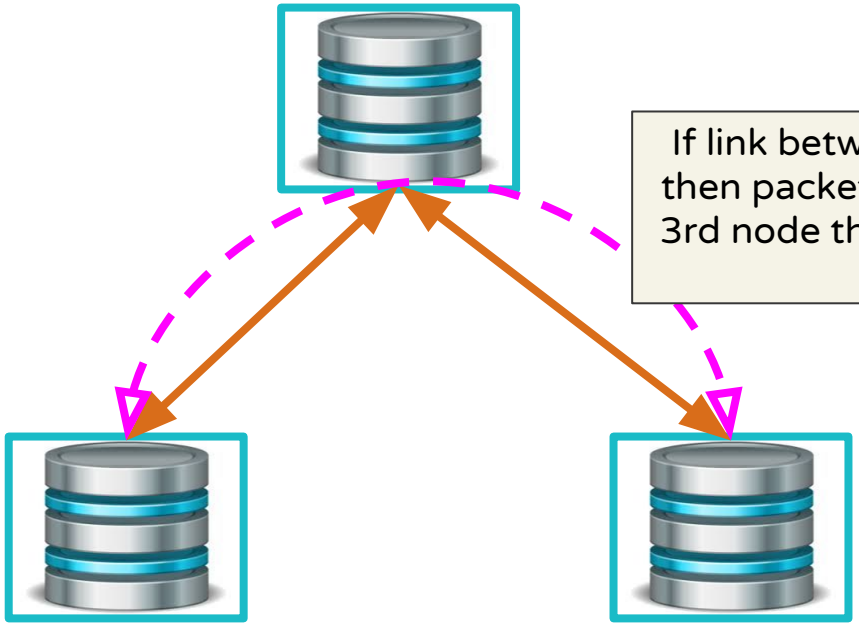
All my writes are going to single node still I am getting this conflict ?

Scenario: Network latency and related failures



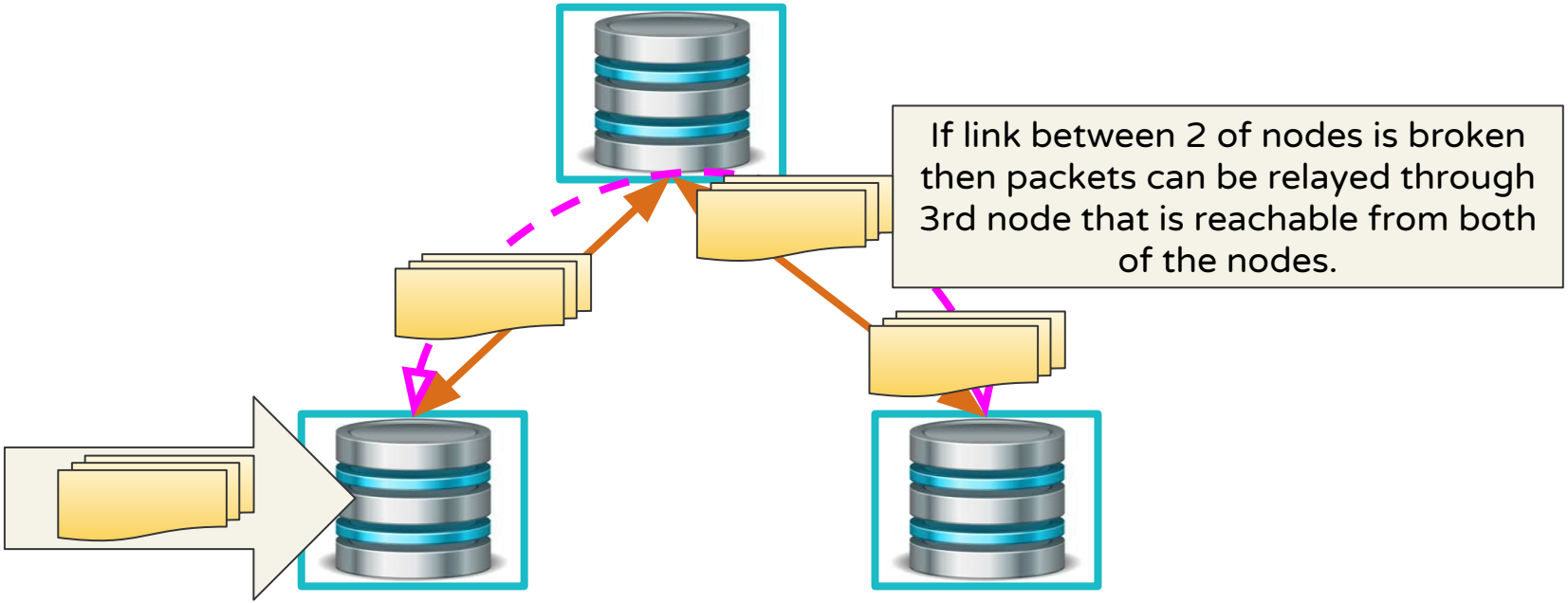
All nodes are able to reach each other

Scenario: Network latency and related failures

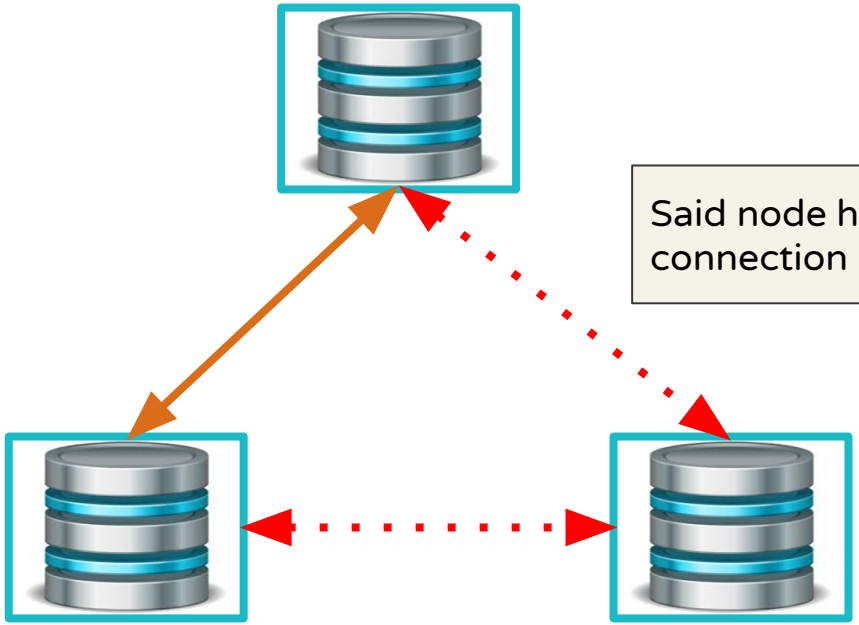


If link between 2 of nodes is broken then packets can be relayed through 3rd node that is reachable from both of the nodes.

Scenario: Network latency and related failures

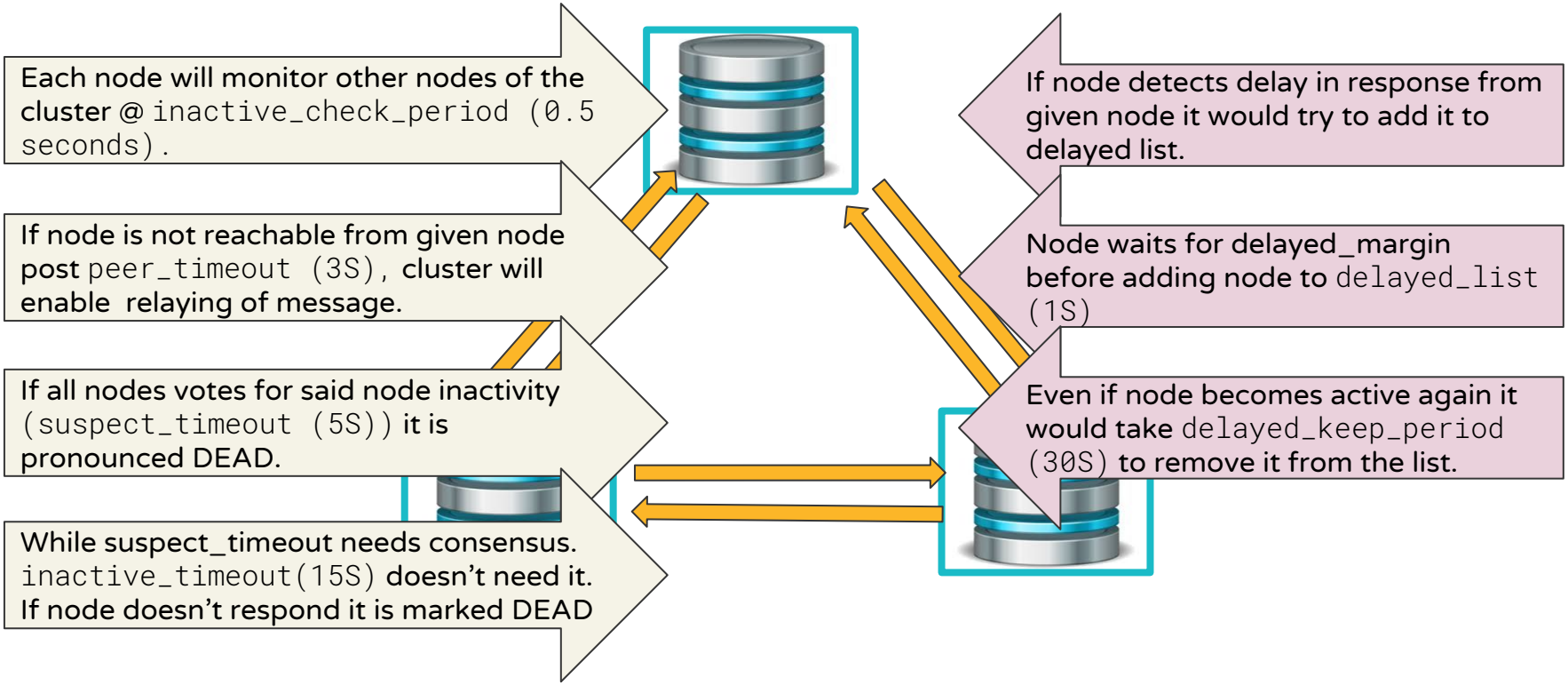


Scenario: Network latency and related failures

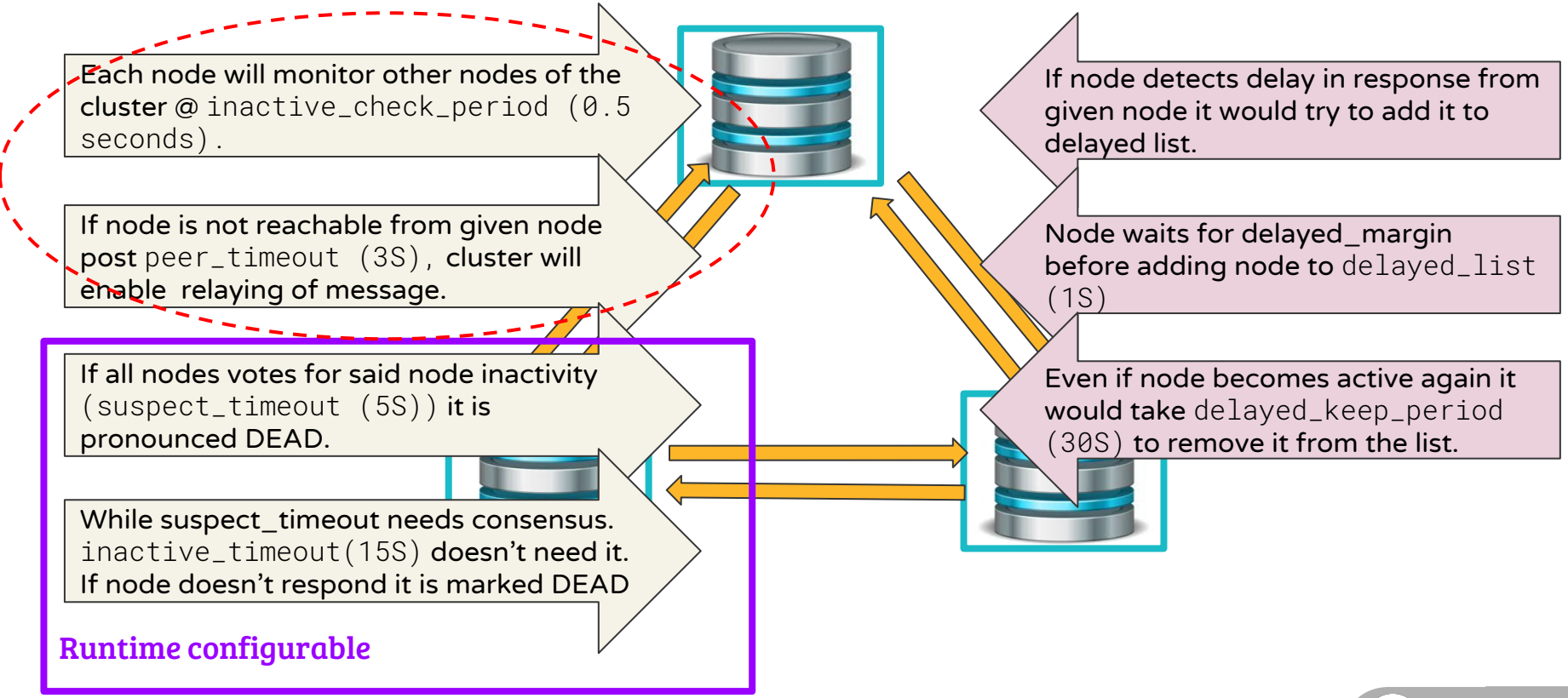


Said node has flaky network connection or say has higher latency.

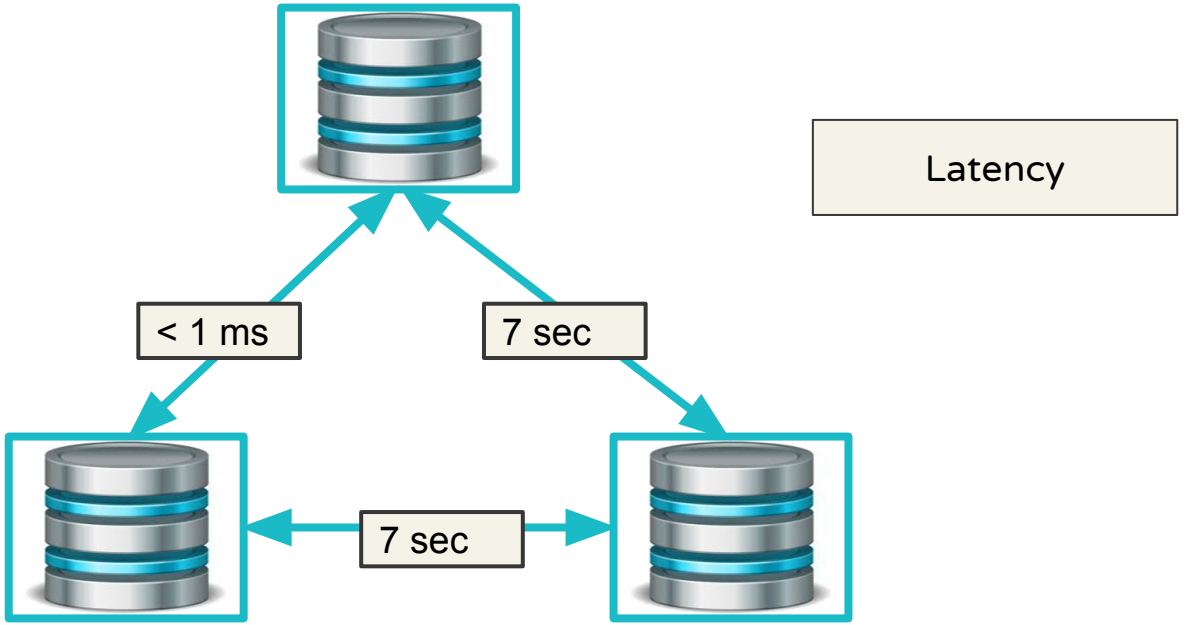
Scenario: Network latency and related failures



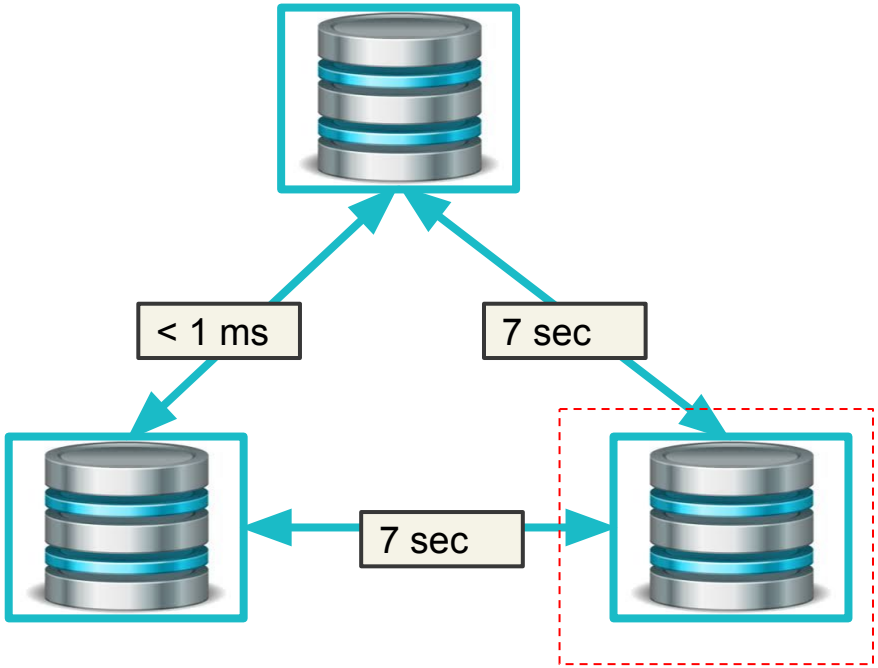
Scenario: Network latency and related failures



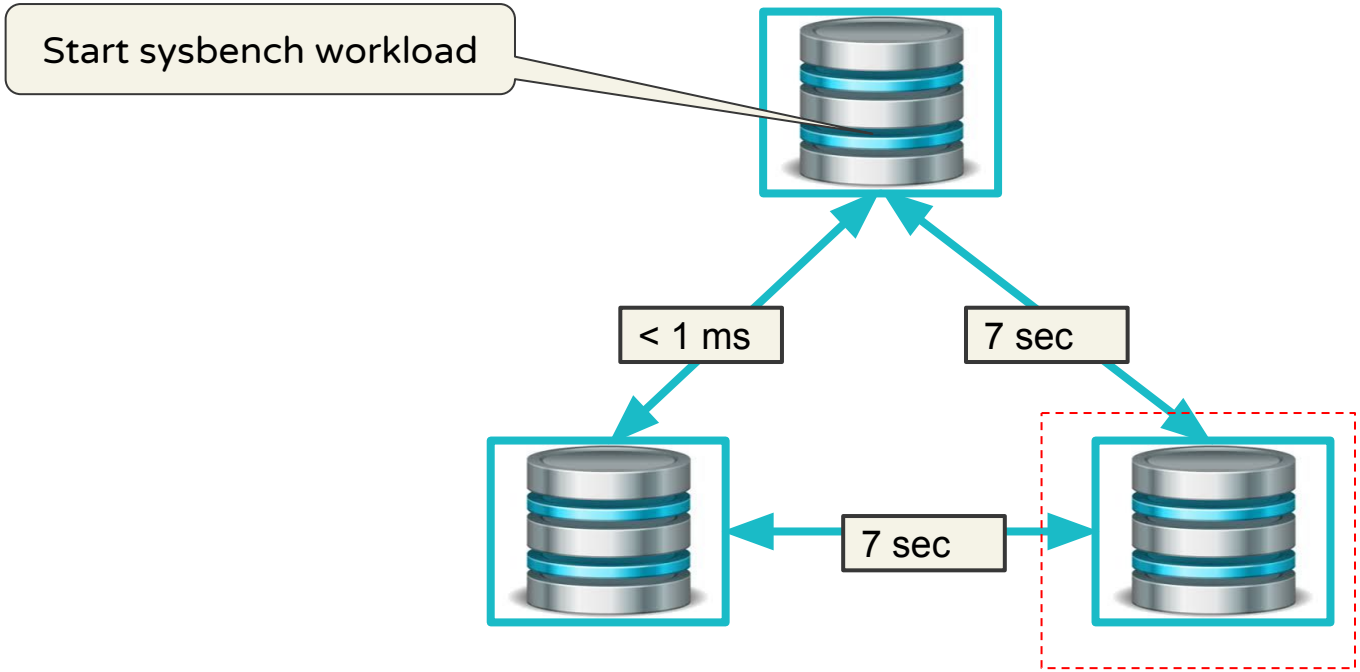
Scenario: Network latency and related failures



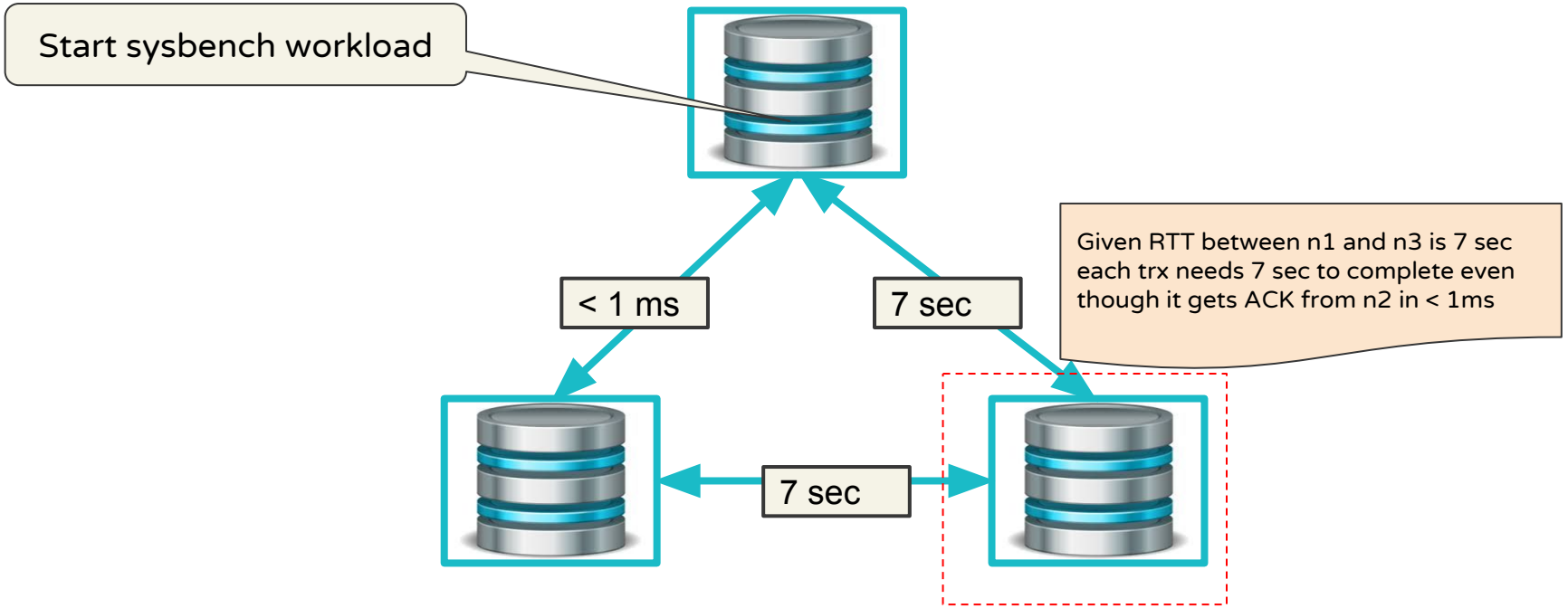
Scenario: Network latency and related failures



Scenario: Network latency and related failures



Scenario: Network latency and related failures



Scenario: Network latency and related failures

```
[ 1s ] thds: 1 tps: 2356.12 qps: 2356.12 (r/w/o: 0.00/751.76/1604.36) lat (ms,99%): 1.58 err/s: 0.00 reconn/s: 0.00
[ 2s ] thds: 1 tps: 2223.44 qps: 2223.44 (r/w/o: 0.00/728.82/1494.63) lat (ms,99%): 1.73 err/s: 0.00 reconn/s: 0.00
[ 3s ] thds: 1 tps: 2100.64 qps: 2100.64 (r/w/o: 0.00/718.22/1382.42) lat (ms,99%): 1.76 err/s: 0.00 reconn/s: 0.00
[ 4s ] thds: 1 tps: 130.39 qps: 130.39 (r/w/o: 0.00/44.45/85.94) lat (ms,99%): 3.49 err/s: 0.00 reconn/s: 0.00
[ 5s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 6s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 7s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 8s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 9s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 10s ] thds: 1 tps: 425.28 qps: 425.28 (r/w/o: 0.00/126.68/298.60) lat (ms,99%): 24.38 err/s: 0.00 reconn/s: 0.00
[ 11s ] thds: 1 tps: 1017.01 qps: 1017.01 (r/w/o: 0.00/358.83/658.17) lat (ms,99%): 10.09 err/s: 0.00 reconn/s: 0.00
[ 12s ] thds: 1 tps: 1668.05 qps: 1668.05 (r/w/o: 0.00/540.66/1127.39) lat (ms,99%): 4.03 err/s: 0.00 reconn/s: 0.00
[ 13s ] thds: 1 tps: 377.62 qps: 377.62 (r/w/o: 0.00/143.85/233.76) lat (ms,99%): 38.25 err/s: 0.00 reconn/s: 0.00
[ 14s ] thds: 1 tps: 2272.97 qps: 2272.97 (r/w/o: 0.00/752.97/1520.00) lat (ms,99%): 1.96 err/s: 0.00 reconn/s: 0.00
[ 15s ] thds: 1 tps: 610.60 qps: 610.60 (r/w/o: 0.00/213.51/397.09) lat (ms,99%): 1.82 err/s: 0.00 reconn/s: 0.00
[ 16s ] thds: 1 tps: 1888.19 qps: 1888.19 (r/w/o: 0.00/632.40/1255.79) lat (ms,99%): 2.14 err/s: 0.00 reconn/s: 0.00
[ 17s ] thds: 1 tps: 2198.33 qps: 2198.33 (r/w/o: 0.00/732.11/1466.22) lat (ms,99%): 1.52 err/s: 0.00 reconn/s: 0.00
[ 18s ] thds: 1 tps: 2143.21 qps: 2143.21 (r/w/o: 0.00/698.74/1444.47) lat (ms,99%): 1.76 err/s: 0.00 reconn/s: 0.00
[ 19s ] thds: 1 tps: 2180.45 qps: 2180.45 (r/w/o: 0.00/735.15/1445.30) lat (ms,99%): 1.55 err/s: 0.00 reconn/s: 0.00
[ 20s ] thds: 1 tps: 2094.64 qps: 2094.64 (r/w/o: 0.00/710.88/1383.76) lat (ms,99%): 1.76 err/s: 0.00 reconn/s: 0.00
[ 21s ] thds: 1 tps: 1668.31 qps: 1668.31 (r/w/o: 0.00/593.11/1075.20) lat (ms,99%): 3.68 err/s: 0.00 reconn/s: 0.00
[ 22s ] thds: 1 tps: 1827.83 qps: 1827.83 (r/w/o: 0.00/647.94/1179.89) lat (ms,99%): 2.48 err/s: 0.00 reconn/s: 0.00
[ 23s ] thds: 1 tps: 2142.13 qps: 2142.13 (r/w/o: 0.00/704.04/1438.09) lat (ms,99%): 1.70 err/s: 0.00 reconn/s: 0.00
[ 24s ] thds: 1 tps: 194.31 qps: 194.31 (r/w/o: 0.00/68.70/125.62) lat (ms,99%): 1.82 err/s: 0.00 reconn/s: 0.00
[ 25s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 26s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 27s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 28s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 29s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
```

#1

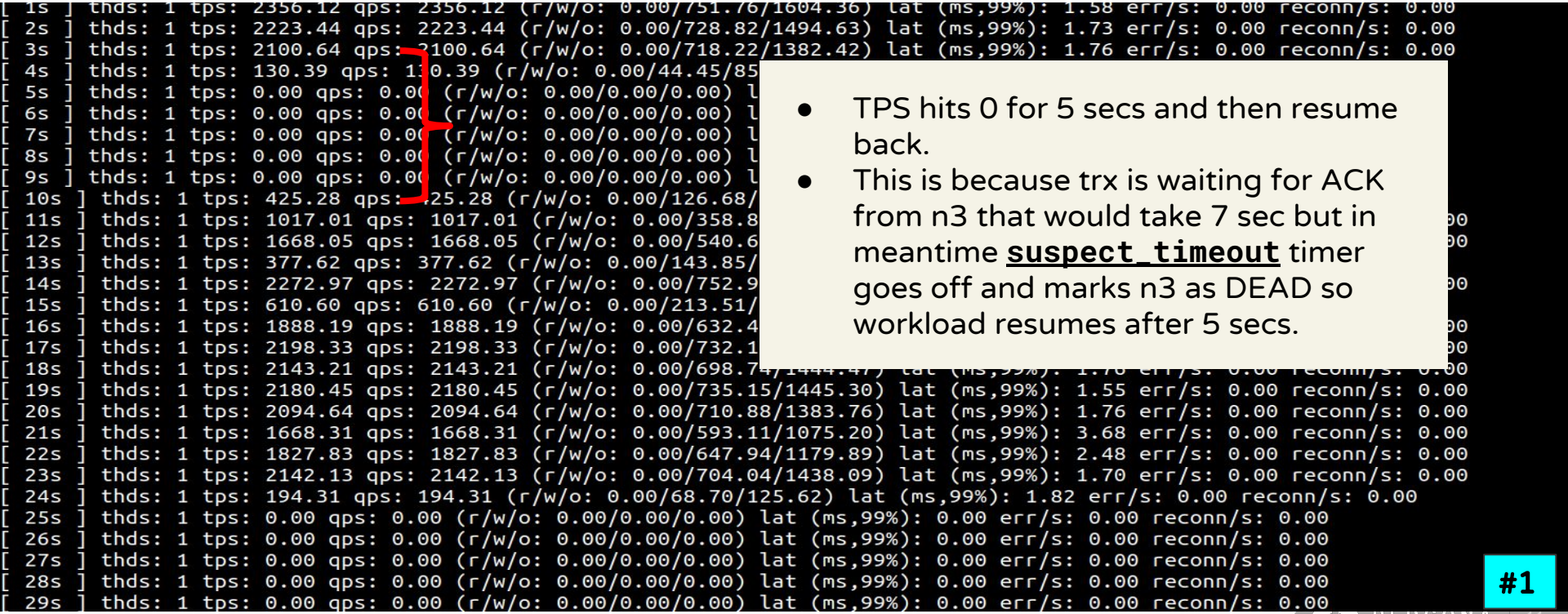
Scenario: Network latency and related failures

```
[ 1s ] thds: 1 tps: 2356.12 qps: 2356.12 (r/w/o: 0.00/751.76/1604.36) lat (ms,99%): 1.58 err/s: 0.00 reconn/s: 0.00
[ 2s ] thds: 1 tps: 2223.44 qps: 2223.44 (r/w/o: 0.00/728.82/1494.63) lat (ms,99%): 1.73 err/s: 0.00 reconn/s: 0.00
[ 3s ] thds: 1 tps: 2100.64 qps: 2100.64 (r/w/o: 0.00/718.22/1382.42) lat (ms,99%): 1.76 err/s: 0.00 reconn/s: 0.00
[ 4s ] thds: 1 tps: 130.39 qps: 130.39 (r/w/o: 0.00/44.45/85
[ 5s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) l
[ 6s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) l
[ 7s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) l
[ 8s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) l
[ 9s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) l
[ 10s ] thds: 1 tps: 425.28 qps: 425.28 (r/w/o: 0.00/126.68/
[ 11s ] thds: 1 tps: 1017.01 qps: 1017.01 (r/w/o: 0.00/358.8
[ 12s ] thds: 1 tps: 1668.05 qps: 1668.05 (r/w/o: 0.00/540.6
[ 13s ] thds: 1 tps: 377.62 qps: 377.62 (r/w/o: 0.00/143.85/
[ 14s ] thds: 1 tps: 2272.97 qps: 2272.97 (r/w/o: 0.00/752.9
[ 15s ] thds: 1 tps: 610.60 qps: 610.60 (r/w/o: 0.00/213.51/
[ 16s ] thds: 1 tps: 1888.19 qps: 1888.19 (r/w/o: 0.00/632.4
[ 17s ] thds: 1 tps: 2198.33 qps: 2198.33 (r/w/o: 0.00/732.1
[ 18s ] thds: 1 tps: 2143.21 qps: 2143.21 (r/w/o: 0.00/698.74/1444.47) lat (ms,99%): 1.70 err/s: 0.00 reconn/s: 0.00
[ 19s ] thds: 1 tps: 2180.45 qps: 2180.45 (r/w/o: 0.00/735.15/1445.30) lat (ms,99%): 1.55 err/s: 0.00 reconn/s: 0.00
[ 20s ] thds: 1 tps: 2094.64 qps: 2094.64 (r/w/o: 0.00/710.88/1383.76) lat (ms,99%): 1.76 err/s: 0.00 reconn/s: 0.00
[ 21s ] thds: 1 tps: 1668.31 qps: 1668.31 (r/w/o: 0.00/593.11/1075.20) lat (ms,99%): 3.68 err/s: 0.00 reconn/s: 0.00
[ 22s ] thds: 1 tps: 1827.83 qps: 1827.83 (r/w/o: 0.00/647.94/1179.89) lat (ms,99%): 2.48 err/s: 0.00 reconn/s: 0.00
[ 23s ] thds: 1 tps: 2142.13 qps: 2142.13 (r/w/o: 0.00/704.04/1438.09) lat (ms,99%): 1.70 err/s: 0.00 reconn/s: 0.00
[ 24s ] thds: 1 tps: 194.31 qps: 194.31 (r/w/o: 0.00/68.70/125.62) lat (ms,99%): 1.82 err/s: 0.00 reconn/s: 0.00
[ 25s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 26s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 27s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 28s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 29s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
```

- TPS hits 0 for 5 secs and then resume back.

#1

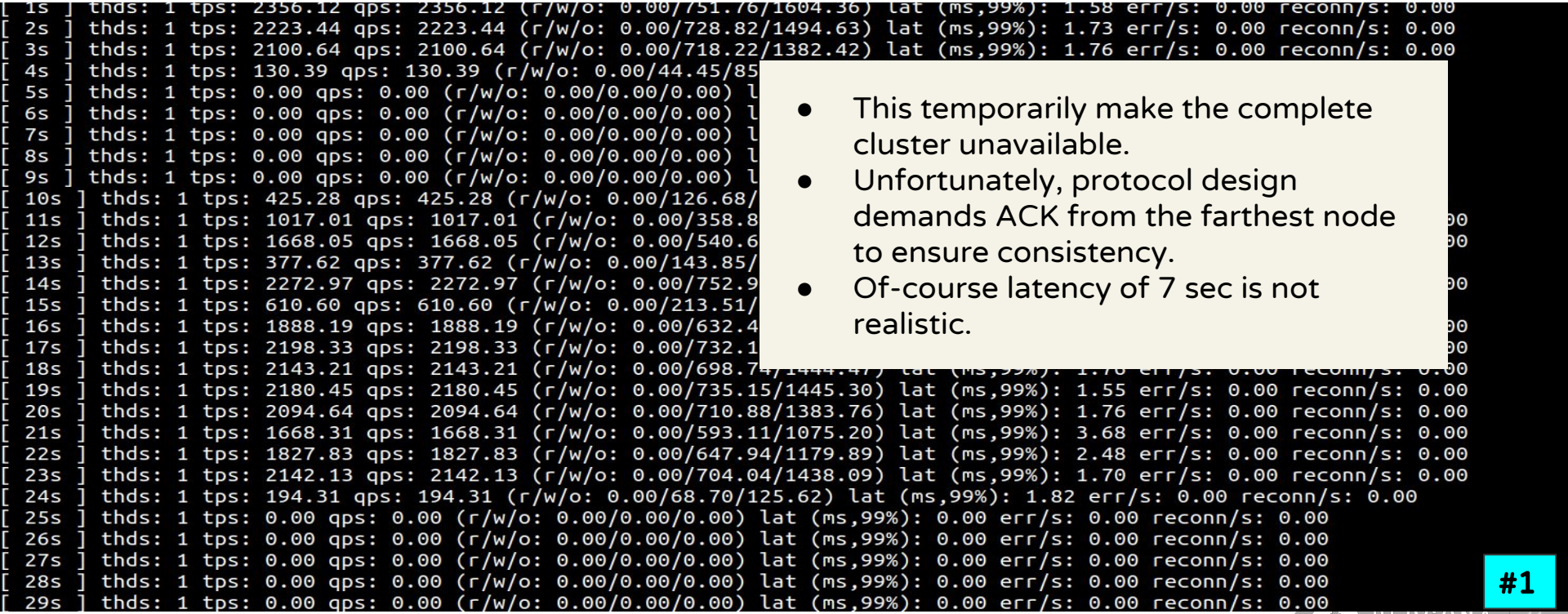
Scenario: Network latency and related failures



- TPS hits 0 for 5 secs and then resume back.
- This is because trx is waiting for ACK from n3 that would take 7 sec but in meantime **`suspect_timeout`** timer goes off and marks n3 as DEAD so workload resumes after 5 secs.

#1

Scenario: Network latency and related failures



- This temporarily make the complete cluster unavailable.
- Unfortunately, protocol design demands ACK from the farthest node to ensure consistency.
- Of-course latency of 7 sec is not realistic.

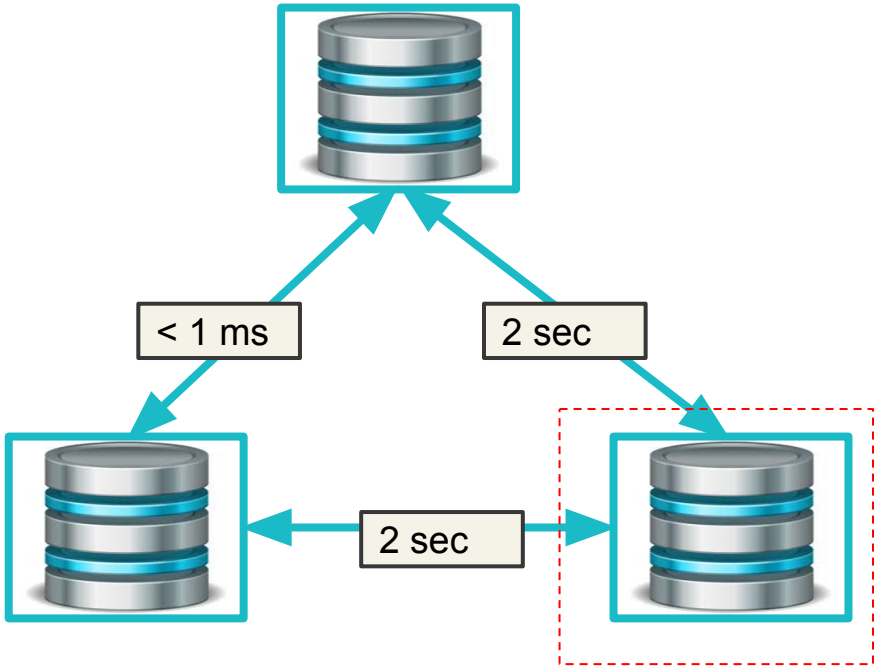
#1

Scenario: Network latency and related failures

```
1s ] thds: 1 tps: 1004.40 qps: 1004.40 (r/w/o: 0.00/330.47/673.93) lat (ms,99%): 2.61 err/s: 0.00 reconn/s: 0.00
2s ] thds: 1 tps: 1798.95 qps: 1798.95 (r/w/o: 0.00/604.98/1193.96) lat (ms,99%): 2.22 err/s: 0.00 reconn/s: 0.00
3s ] thds: 1 tps: 1761.28 qps: 1761.28 (r/w/o: 0.00/546.09/1215.19) lat (ms,99%): 4.49 err/s: 0.00 reconn/s: 0.00
4s ] thds: 1 tps: 2205.29 qps: 2205.29 (r/w/o: 0.00/756.10/1449.19) lat (ms,99%): 1.58 err/s: 0.00 reconn/s: 0.00
5s ] thds: 1 tps: 2256.02 qps: 2256.02 (r/w/o: 0.00/764.01/1492.02) lat (ms,99%): 1.52 err/s: 0.00 reconn/s: 0.00
6s ] thds: 1 tps: 2204.72 qps: 2204.72 (r/w/o: 0.00/749.90/1454.81) lat (ms,99%): 1.64 err/s: 0.00 reconn/s: 0.00
7s ] thds: 1 tps: 2273.16 qps: 2273.16 (r/w/o: 0.00/716.05/1557.11) lat (ms,99%): 1.82 err/s: 0.00 reconn/s: 0.00
8s ] thds: 1 tps: 2232.03 qps: 2232.03 (r/w/o: 0.00/755.01/1477.02) lat (ms,99%): 1.58 err/s: 0.00 reconn/s: 0.00
9s ] thds: 1 tps: 2074.80 qps: 2074.80 (r/w/o: 0.00/701.93/1372.87) lat (ms,99%): 2.11 err/s: 0.00 reconn/s: 0.00
10s ] thds: 1 tps: 186.94 qps: 186.94 (r/w/o: 0.00/58.05/128.89) lat (ms,99%): 2.97 err/s: 0.00 reconn/s: 0.00
11s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
12s ] thds: 1 tps: 2.01 qps: 2.01 (r/w/o: 0.00/1.01/1.01) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
13s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
14s ] thds: 1 tps: 2.93 qps: 2.93 (r/w/o: 0.00/0.98/1.95) lat (ms,99%): 2320.55 err/s: 0.00 reconn/s: 0.00
15s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
16s ] thds: 1 tps: 1.00 qps: 1.00 (r/w/o: 0.00/1.00/0.00) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
17s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
18s ] thds: 1 tps: 1.00 qps: 1.00 (r/w/o: 0.00/1.00/0.00) lat (ms,99%): 2009.23 err/s: 0.00 reconn/s: 0.00
19s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
20s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
21s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
22s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
23s ] thds: 1 tps: 1384.62 qps: 1384.62 (r/w/o: 0.00/472.99/911.62) lat (ms,99%): 2.26 err/s: 0.00 reconn/s: 0.00
24s ] thds: 1 tps: 2224.98 qps: 2224.98 (r/w/o: 0.00/719.29/1505.69) lat (ms,99%): 2.07 err/s: 0.00 reconn/s: 0.00
25s ] thds: 1 tps: 2061.52 qps: 2061.52 (r/w/o: 0.00/690.84/1370.68) lat (ms,99%): 1.86 err/s: 0.00 reconn/s: 0.00
26s ] thds: 1 tps: 1814.65 qps: 1814.65 (r/w/o: 0.00/590.21/1224.44) lat (ms,99%): 2.22 err/s: 0.00 reconn/s: 0.00
27s ] thds: 1 tps: 1794.67 qps: 1794.67 (r/w/o: 0.00/587.89/1206.78) lat (ms,99%): 2.18 err/s: 0.00 reconn/s:
28s ] thds: 1 tps: 2086.06 qps: 2086.06 (r/w/o: 0.00/669.02/1417.04) lat (ms,99%): 1.67 err/s: 0.00 reconn/s:
```

#2

Scenario: Network latency and related failures



Scenario: Network latency and related failures

```
1s ] thds: 1 tps: 1004.40 qps: 1004.40 (r/w/o: 0.00/330.47/673.93) lat (ms,99%): 2.61 err/s: 0.00 reconn/s: 0.00
2s ] thds: 1 tps: 1798.95 qps: 1798.95 (r/w/o: 0.00/604.98/1193.96) lat (ms,99%): 2.22 err/s: 0.00 reconn/s: 0.00
3s ] thds: 1 tps: 1761.28 qps: 1761.28 (r/w/o: 0.00/546.09/1215.19) lat (ms,99%): 4.49 err/s: 0.00 reconn/s: 0.00
4s ] thds: 1 tps: 2205.29 qps: 2205.29 (r/w/o: 0.00/756.10/1449.19) lat (ms,99%): 1.58 err/s: 0.00 reconn/s: 0.00
5s ] thds: 1 tps: 2256.02 qps: 2256.02 (r/w/o: 0.00/764.01
6s ] thds: 1 tps: 2204.72 qps: 2204.72 (r/w/o: 0.00/749.90
7s ] thds: 1 tps: 2273.16 qps: 2273.16 (r/w/o: 0.00/716.05
8s ] thds: 1 tps: 2232.03 qps: 2232.03 (r/w/o: 0.00/755.01
9s ] thds: 1 tps: 2074.80 qps: 2074.80 (r/w/o: 0.00/701.93
10s ] thds: 1 tps: 186.94 qps: 186.94 (r/w/o: 0.00/58.05/1
11s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00)
12s ] thds: 1 tps: 2.01 qps: 2.01 (r/w/o: 0.00/1.01/1.01)
13s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00)
14s ] thds: 1 tps: 2.93 qps: 2.93 (r/w/o: 0.00/0.98/1.95)
15s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00)
16s ] thds: 1 tps: 1.00 qps: 1.00 (r/w/o: 0.00/1.00/0.00)
17s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00)
18s ] thds: 1 tps: 1.00 qps: 1.00 (r/w/o: 0.00/1.00/0.00)
19s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
20s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
21s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
22s ] thds: 1 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
23s ] thds: 1 tps: 1384.62 qps: 1384.62 (r/w/o: 0.00/472.99/911.62) lat (ms,99%): 2.26 err/s: 0.00 reconn/s: 0.00
24s ] thds: 1 tps: 2224.98 qps: 2224.98 (r/w/o: 0.00/719.29/1505.69) lat (ms,99%): 2.07 err/s: 0.00 reconn/s: 0.00
25s ] thds: 1 tps: 2061.52 qps: 2061.52 (r/w/o: 0.00/690.84/1370.68) lat (ms,99%): 1.86 err/s: 0.00 reconn/s: 0.00
26s ] thds: 1 tps: 1814.65 qps: 1814.65 (r/w/o: 0.00/590.21/1224.44) lat (ms,99%): 2.22 err/s: 0.00 reconn/s: 0.00
27s ] thds: 1 tps: 1794.67 qps: 1794.67 (r/w/o: 0.00/587.89/1206.78) lat (ms,99%): 2.18 err/s: 0.00 reconn/s:
28s ] thds: 1 tps: 2086.06 qps: 2086.06 (r/w/o: 0.00/669.02/1417.04) lat (ms,99%): 1.67 err/s: 0.00 reconn/s:
```

- This time I reduced the latency from 7 to 2 sec. Because of this every 2 sec (less 5 sec) there was some communication between node and this prevent n3 from being marked as DEAD.
- Post 10 secs we reverted back latency to original value so snag is seen for 10 secs.

#2

Scenario: Network latency and related failures

```
2018-10-22T05:28:49.380284Z 7 [Note] WSREP: New cluster view: global state: c0ef3ed5-d1c2-11e8-b7b3-77974da4d118:1091590, view# 14: Primary, number of no
des: 3, my index: 0, protocol version 3
2018-10-22T05:28:49.380287Z 7 [Note] WSREP: Setting wsrep_ready to true
2018-10-22T05:28:49.380289Z 7 [Note] WSREP: Auto Increment Offset/Increment re-align with cluster membership change (Offset: 1 -> 1) (Increment: 3 -> 3)
2018-10-22T05:28:49.380292Z 7 [Note] WSREP: wsrep_notify_cmd is not defined, skipping notification.
2018-10-22T05:28:49.380400Z 7 [Note] WSREP: Assign initial position for certification: 1091590, protocol version: 4
2018-10-22T05:28:49.380553Z 18 [Note] WSREP: --------- CONFLICT DETECTED --------
2018-10-22T05:28:49.380563Z 18 [Note] WSREP: cluster conflict due to certification failure for threads:

2018-10-22T05:28:49.380566Z 18 [Note] WSREP: Victim thread:
   THD: 18, mode: local, state: executing, conflict: cert failure, seqno: -1
   SQL: UPDATE sbtest4 SET c=? WHERE id=?

2018-10-22T05:28:49.380864Z 18 [Note] WSREP: --------- CONFLICT DETECTED --------
2018-10-22T05:28:49.380930Z 18 [Note] WSREP: cluster conflict due to certification failure for threads:

2018-10-22T05:28:49.380934Z 18 [Note] WSREP: Victim thread:
   THD: 18, mode: local, state: executing, conflict: cert failure, seqno: -1
   SQL: UPDATE sbtest2 SET c=? WHERE id=?

#3  22T05:28:49.381183Z 0 [Note] WSREP: Service thread queue flushed.
```

All my writes are going to single node still I am getting this conflict ?

Scenario: Network latency and related failures

```
2018-10-22T05:28:49.380284Z 7 [Note] WSREP: New cluster view: global state: c0ef3ed5-d1c2-11e8-b7b3-77974da4d118:1091590, view# 14: Primary, number of no
des: 3, my index: 0, protocol version 3
2018-10-22T05:28:49.380287Z 7 [Note] WSREP: Setting wsrep_ready to true
2018-10-22T05:28:49.380289Z 7 [Note] WSREP: Auto Increment Offset/Increment re-align with cluster membership change (Offset: 1 -> 1) (Increment: 3 -> 3)
2018-10-22T05:28:49.380292Z 7 [Note] WSREP: wsrep_notify_cmd is not defined, skipping notification.
2018-10-22T05:28:49.380400Z 7 [Note] WSREP: Assign initial position for certification: 1091590, protocol version: 4
2018-10-22T05:28:49.380553Z 18 [Note] WSREP:           CONFLICT DETECTED
2018-10-22T05:28:49.380563Z 18 [Note] WSREP: cluster conflict due to certification failure for threads:

2018-10-22T05:28:49.380566Z 18 [Note] WSREP: Victim thread:
   THD: 18, mode: local, state: executing, conflict: cert failure, seqno: -1
   SQL: UPDATE sbtest4 SET c=? WHERE id=?

2018-10-22T05:28:49.380864Z 18 [Note] WSREP: --------- CONFLICT DETECTED --------
2018-10-22T05:28:49.380930Z 18 [Note] WSREP: cluster conflict due to certification

2018-10-22T05:28:49.380934Z 18 [Note] WSREP: Victim thread:
   THD: 18, mode: local, state: executing, conflict: cert failure, seqno: -1
   SQL: UPDATE sbtest2 SET c=? WHERE id=?

22T05:28:49.381183Z 0 [Note] WSREP: Service thread queue flushed.
```

#3

Because when the view changes initial position is re-assigned there-by purging history from cert index. Follow up transaction in cert that has dependency with old trx (that got purged) faces this conflict.

Scenario: Network latency and related failures

Farthest node dictates how cluster would operate and so latency is important.

Geo-Distributed cluster has milli-sec latency so timeout should be configured to avoid marking node as UNSTABLE due to added latency.

For geo-distributed cluster segment, window settings are other param to configure.

Flaky node are not good for overall transaction processing. (Can cause certification failures).

Scenario: Blocking Transaction and related failures

Scenario: Blocking Transaction and related failures

```
mysql> create table tmp like sbtest1;
Query OK, 0 rows affected (0.03 sec)

mysql> insert into tmp select * from sbtest1;
ERROR 1180 (HY000): Got error 5 during COMMIT
mysql>
```

- Fail to load a table with N rows.

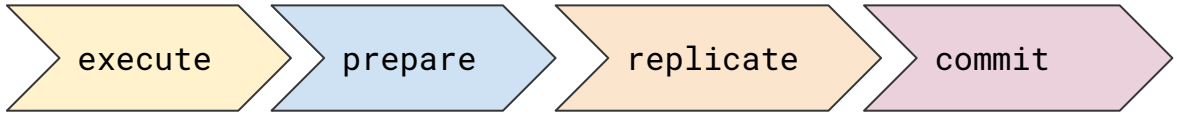
Scenario: Blocking Transaction and related failures

```
mysql> create table tmp like sbtest1;
Query OK, 0 rows affected (0.03 sec)

mysql> insert into tmp select * from sbtest1;
ERROR 1180 (HY000): Got error 5 during COMMIT
mysql>
```

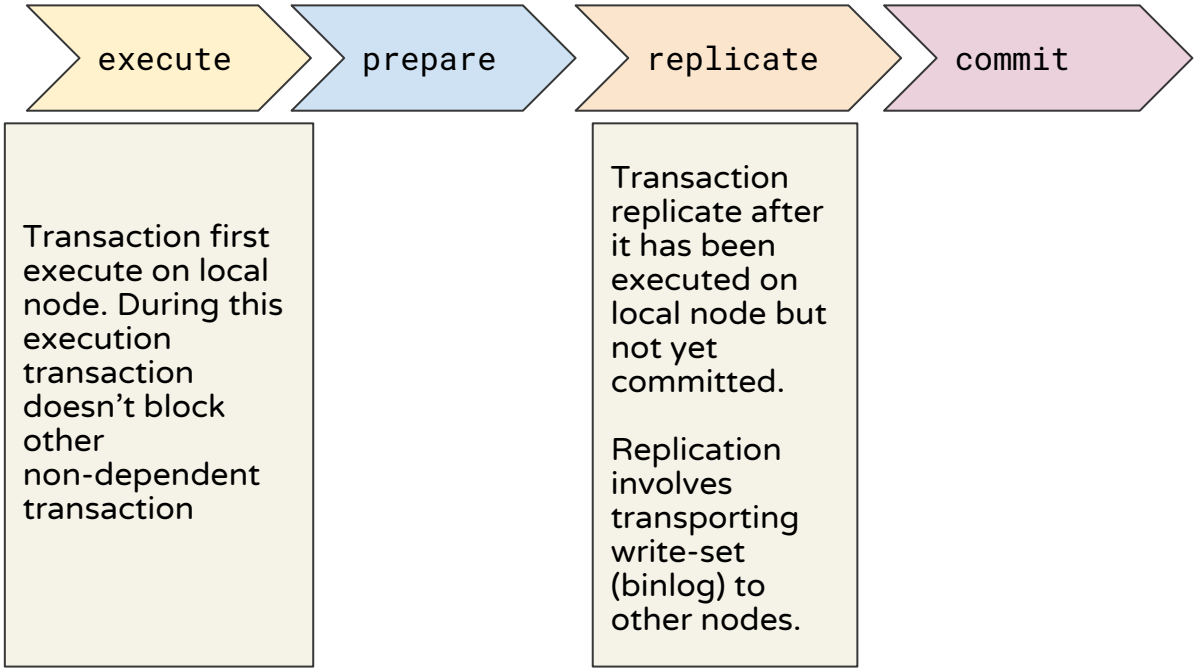
- Fail to load a table with N rows.
- Why ?
  - Because PXC has limit on how much data it can wrap in write-set and replicate across the cluster.
  - Current limit allows data transaction of size 2 G. (controlled through wsrep_max_ws_size)

But ever imagined why is that a limitation ?

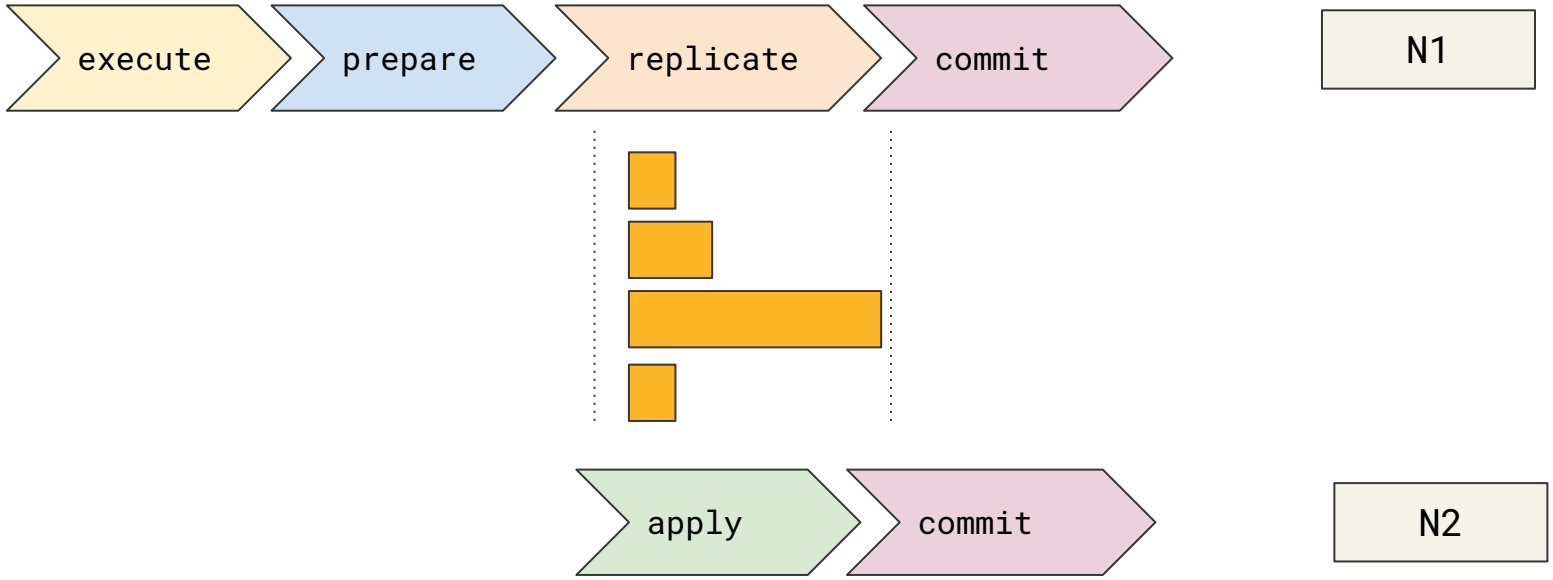
Scenario: Blocking Transaction and related failures



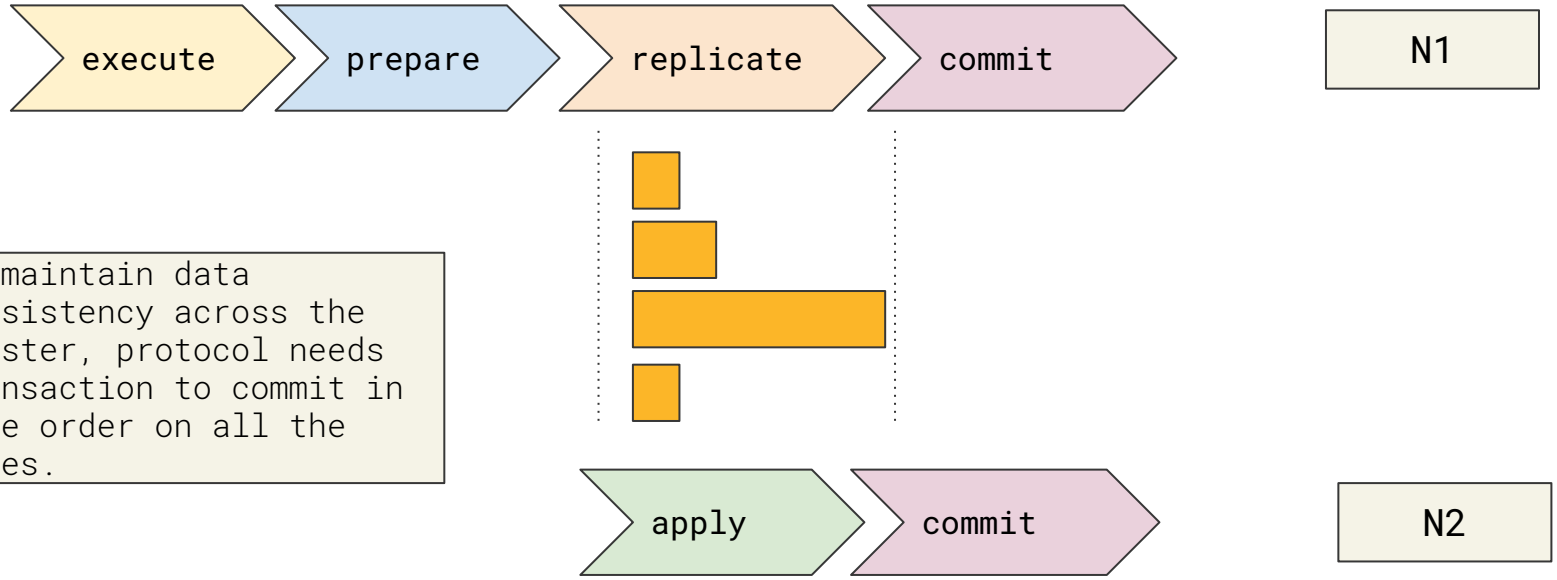
Scenario: Blocking Transaction and related failures



Scenario: Blocking Transaction and related failures

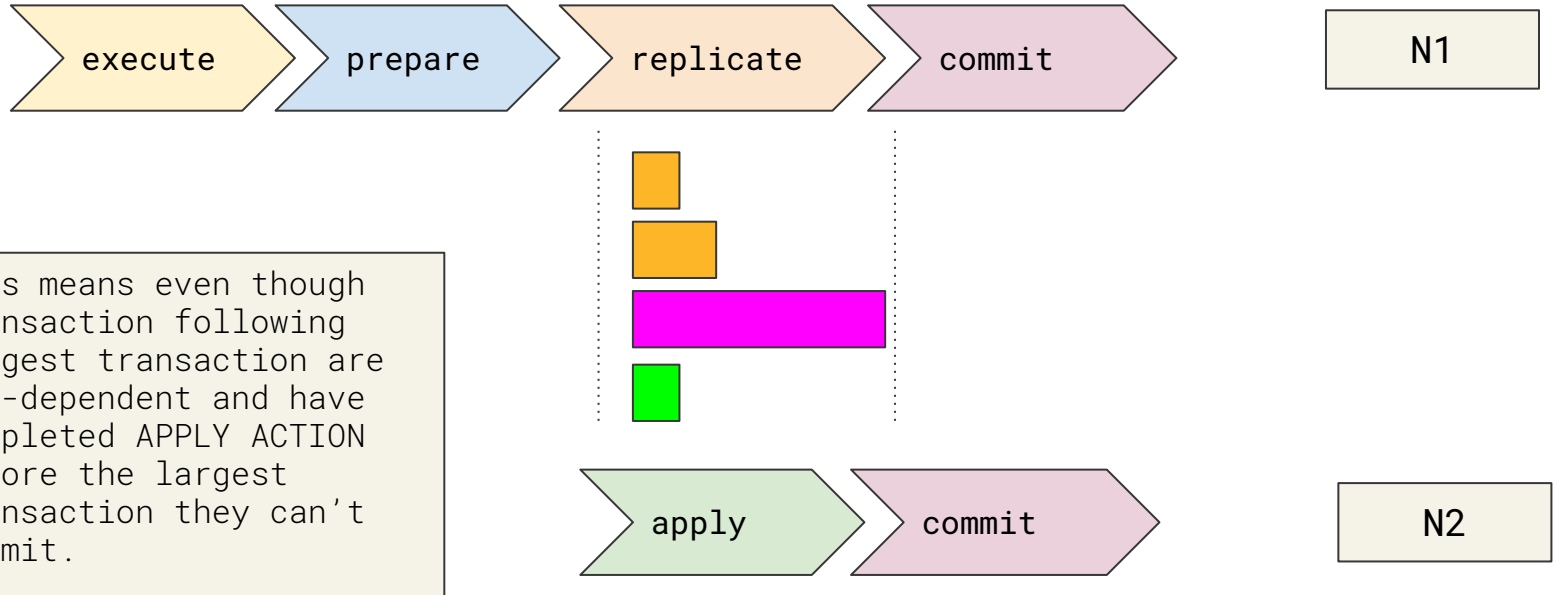


Scenario: Blocking Transaction and related failures



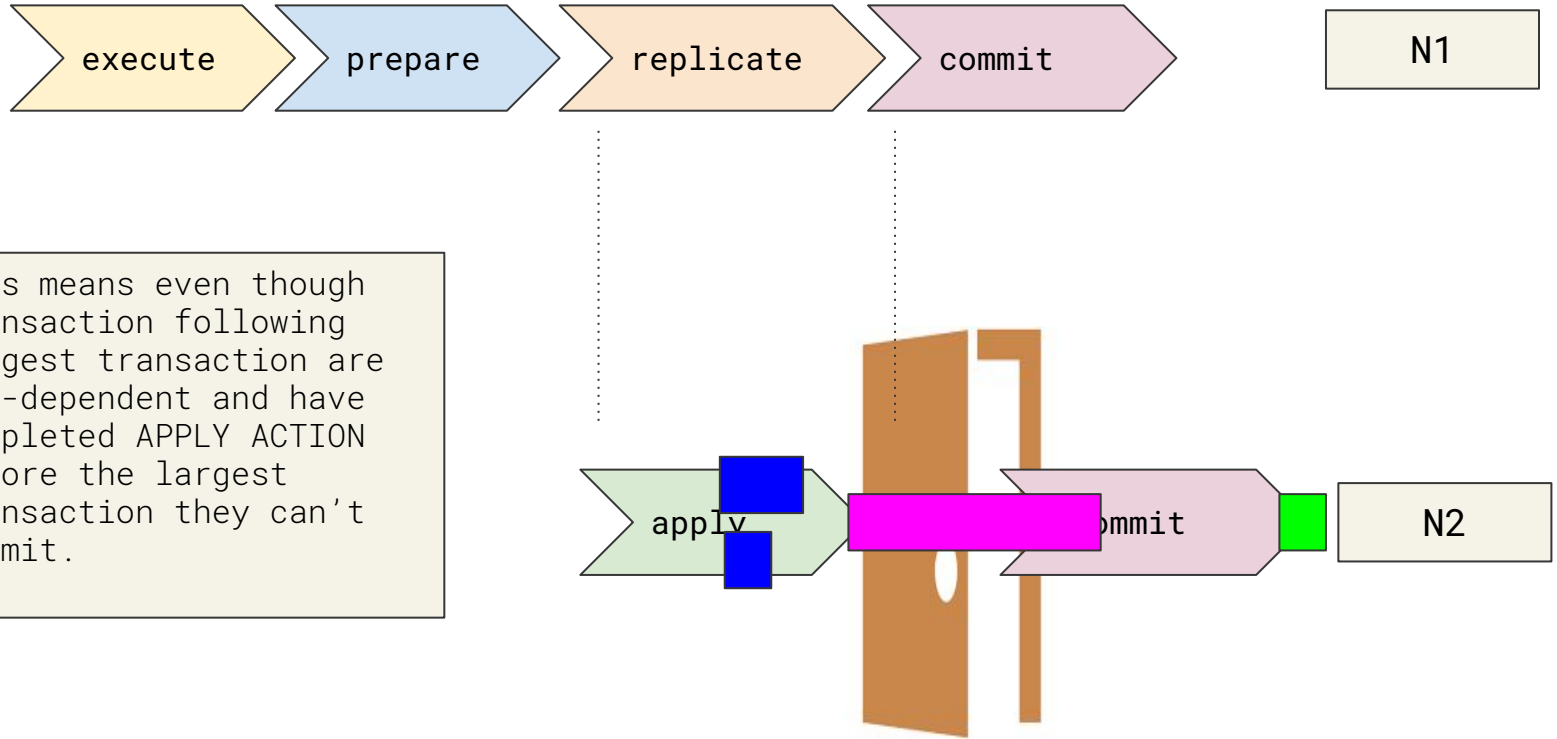
To maintain data consistency across the cluster, protocol needs transaction to commit in same order on all the nodes.

Scenario: Blocking Transaction and related failures



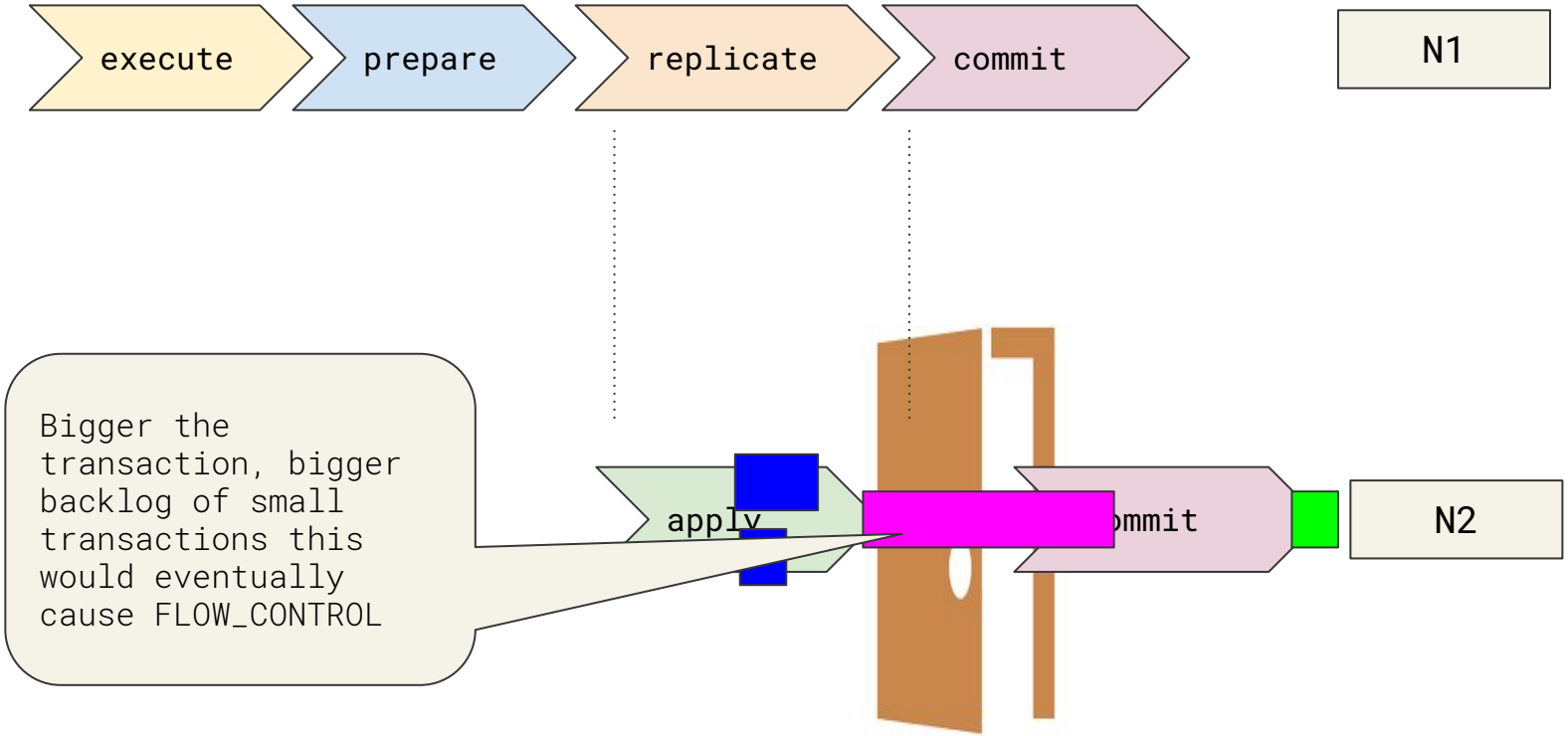
This means even though transaction following largest transaction are non-dependent and have completed APPLY ACTION before the largest transaction they can't commit.

Scenario: Blocking Transaction and related failures



This means even though transaction following largest transaction are non-dependent and have completed APPLY ACTION before the largest transaction they can't commit.

Scenario: Blocking Transaction and related failures



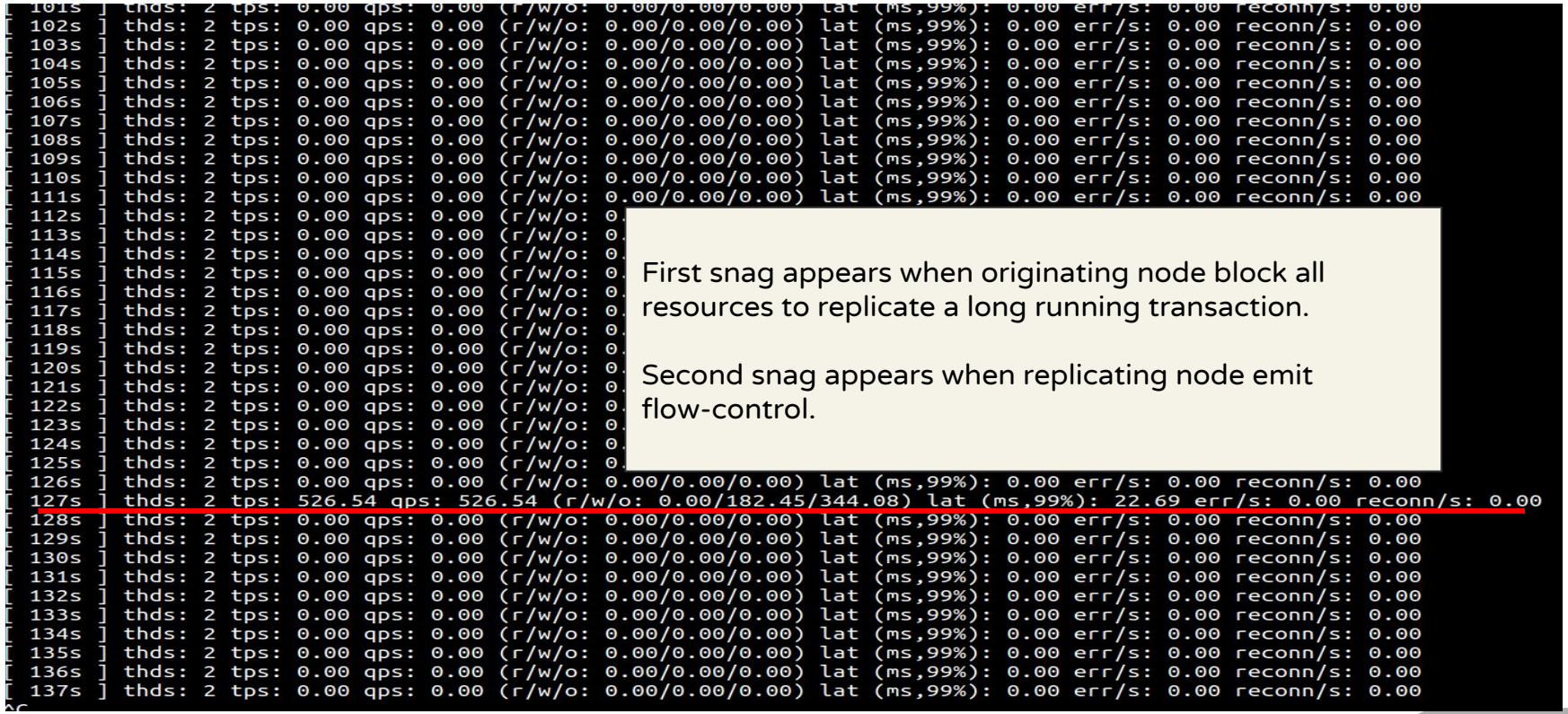
Scenario: Blocking Transaction and related failures

```
[ 101s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 102s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 103s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 104s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 105s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 106s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 107s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 108s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 109s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 110s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 111s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 112s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 113s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 114s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 115s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 116s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 117s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 118s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 119s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 120s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 121s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 122s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 123s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 124s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 125s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 126s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 127s ] thds: 2 tps: 526.54 qps: 526.54 (r/w/o: 0.00/182.45/344.08) lat (ms,99%): 22.69 err/s: 0.00 reconn/s: 0.00
[ 128s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 129s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 130s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 131s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 132s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 133s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 134s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 135s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 136s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 137s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
^C
```

Scenario: Blocking Transaction and related failures

```
[ 101s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 102s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 103s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 104s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 105s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 106s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 107s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 108s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 109s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 110s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 111s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 112s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 113s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 114s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 115s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 116s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 117s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 118s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 119s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 120s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 121s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 122s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 123s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 124s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 125s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 126s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 127s ] thds: 2 tps: 526.54 qps: 526.54 (r/w/o: 0.00/182.45/344.08) lat (ms,99%): 22.69 err/s: 0.00 reconn/s: 0.00
[ 128s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 129s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 130s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 131s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 132s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 133s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 134s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 135s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 136s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
[ 137s ] thds: 2 tps: 0.00 qps: 0.00 (r/w/o: 0.00/0.00/0.00) lat (ms,99%): 0.00 err/s: 0.00 reconn/s: 0.00
^C
```

Scenario: Blocking Transaction and related failures



First snag appears when originating node block all resources to replicate a long running transaction.

Second snag appears when replicating node emit flow-control.

Scenario: Network latency and related failures

PXC doesn't like long running transaction.

For load data use LOAD DATA INFILE that would cause intermediate commit every 10K rows. *Note: Random failure can cause partial data to get committed.*

DDL can block/stall complete cluster workload as they need to execute in total-isolation. (Alternative is to use RSU but be careful at it is local operation to the node).

One last important note

- Majority of the error are due to mis-configuration or difference in configuration of nodes.
- PXC recommend same configuration on all nodes of the cluster.

PXC Genie: You Wish. We implement

PXC Genie: You Wish. We implement

- Like to hear from you what you want next in PXC ?
- Any specific module that you expect improvement ?
- How can Percona help you with PXC or HA ?
- Log issue (mark them as new improvement) https://jira.percona.com/projects/PXC/issue
- PXC forum is other way to reach us.

Questions and Answer